TESIS DOCTORAL

2017

KNOWLEDGE-BASED CATALOGUING OF DEEP MULTI-WAVELENGTH COSMOLOGICAL SURVEYS

by

María José Márquez Sánchez

PROGRAMA DE DOCTORADO EN SISTEMAS INTELIGENTES

PhD Advisor: Dr. Sarro Baro, Luis Manuel. UNED University.

PhD Co-Advisor: Dr. Budavári, Tamás. JHU University .

A dissertation submitted to UNED University in conformity with the requirements for the degree of Doctor of Philosophy.

Madrid, Spain

July, 2017

1. Reviewer: Name

2. Reviewer:

Day of the defense:

Signature from head of PhD committee:

Abstract

I developed an expert data processing pipeline to improve the quality of the astronomical multi-colour galaxy catalogues. This pipeline acts in three main areas: (1) refining surface brightness measurements, (2) flagging contaminated sources, and (3) improving the crossmatching of galaxies by modelling their colours. In order to develop this expert system, we start by building the knowledge model of entities and tasks involved and, as the result of this, we populate an ontology with instances of these classes. This ontology was created in line with the existing ones of the International Virtual Observatory Alliance(IVOA) initiative which intends to unify the Astronomy data models used across the scientific community.

We achieved this goal through the use of artificial intelligence methodologies. We used Bayesian model decision, active contour from Artificial Vision and rule-based systems, all three of these methodologies are integrated into a single modular expert system versatile and flexible enough to be used partially or entirely as a module of another processing system, as well as an input for a refined IVOA ontology. We performed a validation of the system with synthetic data and real astronomical images

and catalogues from the Cosmic Evolution Survey (COSMOS). The outcome of this step confirmed the validity of our proposal. The results on a reduced but representative COSMOS data set confirmed that the photometric cross matching solution and the labelling of sources to determine their contamination status is a significant improvement with respect to the existing methodologies. Regarding the refinement of source contours, the experiment conducted with synthetic data as well as the results on real data demonstrated that this methodology is more flexible in the determination of contours and it allows a good degree of automatism; however the computational cost and the observed lack of certainty for all cases indicates that this proposal needs further development. Overall it can be concluded that our system constitutes a significant advance in the cataloguing of galaxies in multi-band image cubes. We achieved this with an expert system that can be run without much human supervision.

Acknowledgments

I cannot sufficiently express in words my gratefulness and the admiration I feel towards my parents who taught me how to enjoy learning since my very first day of life; they showed me how to explore the path of understanding in life, without prejudgement. They explained to me how to fly with my soul to touch all that life gives me and also all that is taken away from this world.

I would like to acknowledge a profound gratitude towards my sister Salve and brother Francisco, the best possible supports that one can think of in the path of life. Both inspirational sources of my goals. Sharing the core of the same legacy and school has given me strength in difficult moments.

I would like to dedicate here some special words of thanks to my first academic mentor and very good friend, Dr. Francisco Javier Ríos Gomez, who not only showed me how extraordinary is to feel passionate about Science but also discovered my potential in Science before I was aware of it.

Thanks to my PhD advisor, Dr. Luis Manuel Sarro Baro who encouraged me towards the path of this research and let me walk through it while standing on my own feet,

who also came to my assistance every time I faced seemingly insurmountable difficulties. His excellent advice and guidance made me grow in strength and wisdom while constantly stimulating my passion for Science. I also acknowledge here a profound admiration for his professional and human integrity - shining examples for me to follow.

It is a privilege to acknowledge with deep gratefulness, Dr. Tamás Budavári. I could not have done this research without his absolutely excellent support. I am extremely lucky to have been able to count on his advice and guidance allowing me to learn from an incredibly brilliant mind and an extraordinary professor. His advice in this research has been invaluable to me and cultivated in me even more motivation in this research. The time I spent in Johns Hopkins University in a very kind welcoming context under his guidance, was the most scientifically insightful and efficient one in terms of learning that I have ever had. A Big thank you for that.

I would like to mention here with deep gratefulness and appreciation Dr. Felix de la Paz, for his crucial support in one of the most challenging moments of this research.

I acknowledge here with profound gratefulness the support provided by my work organisation, EUMETSAT. This community, to which I belong and feel proud of, has provided me with an extraordinary support and good understanding of my motivation in Science. From the Director General, Mr. Alain Ratier going through my hierarchical management chain, Mr. Yves Buhler, Mr. Lorenzo Sarlo, Mr. Gianfranco Rimoldi, Mr. Chris Hartley, they represent to me a solid and sound reference in my

path and I acknowledge here with profound gratefulness that without their support this work would not have been possible.

Thanks to the colleagues of my team (GEO QA team), as well as colleagues from other departments and hierarchical levels, such as Mr. Kenneth Holmlund, Mr. Jochen Grandel, Mr. Stephen Tjemkes, Ms. Marie Doutriaux Boucher, Mr.Rgis Borde, Mr. Dieter Just, Mr. Daniel Risquez Oneca, Mr. Peter Albert, Mr. Gilles Garnier, Mr. Julian Wilson and all colleagues I met in the Science department, for their insightful discussions in Science and their interest in my research; Ms. Sonia Misouri and the EUMETSAT Astronomy Club, Ms. Margaret Mahler, Ms. Elisabetta Guzzi, Ms. Ester Rojo and, Ms. Beatriz Mora, for their words of motivation towards my research. Special thanks to Ms. Fiona Brazil, Ms. Elayne Chapman and Ms. Silvia Castañer for their effective support. They showed an outstanding level of professionalism and integrity and set up for me an example to follow.

Special thanks to Mr. Georges Bernède. It was a privilege to learn from his professional school during my first years at EUMETSAT. All his wise insights and mentoring helped me in pursuing this research while working in a non-related context. His intelligent approach to problem has always been a inspirational source in my research. Also Big Thanks for his following up of my research and sound corrections in all documentation that I produced.

Big Thanks to Mr. Henry Walls for his generosity in performing thorough and sound English proof reading of the manuscript, as well as for all his wise advice at all times,

being undoubtly a "solid rock of support".

Special thanks to EUMETSAT Technical Computing Environment support team, with special mention to Mr. Ignacio Gonzalez and the Information Communication Technology department, with Mr. Sven Wiessman and his team, comprised by experts such as Mr. Walter Will Aguilar and all other experts of the department. They helped me at all moments with deep and professional expertise in the area of computer science. This support allowed me to work effectively in my research.

To my colleague and friend Ms. Rebeca Martinez, in all honesty, I cannot and will not acknowledge gratefulness in pursuing this research, in view of all her attempts to distract me from it. However, I must also recognize how positive her distraction influence impacted me by my reaction against it and for keeping me more conscious of the importance of using my time better, shared between work and this research. All my sincere indirect gratefulness.

Thanks to Ms. Esperanza Garía Rodríguez, friend of my family who always expressed proud of my accomplishments and motivated me in pursuing new challenges.

Last but not least I acknowledge deep gratefulness to the Jesuit Community in Málaga for their extraordinary support during difficult times. This allowed me to find strength towards the completion of this research. I would like to make special mention here of Mr. Antonio Cobos and Mr. Juan Carlos Hidalgo López.

To all other names not cited here and with whom I shared this extraordinary experience of researching, Big Thanks to each and every one of them.

ACKNOWLEDGMENTS

Credits: This research has made use of the NASA/ IPAC Infrared Science Archive, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

Dedication

This thesis is dedicated to my beloved parents, corner stones of my very existence in this world. All I was, all I am, and all I will be and achieve in my life is thanks to them. They are unfortunately not here to witness their achievement; I therefore present today with honour their legacy through this academic exercise. It is also my wish to devote all my effort in keeping alive in me their schooling. In this sense, the conclusion of this research exercise is the opening of a new door, to be walked through in the spiritual company of my parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

As indicated in [Schneider, 2014], aspects related to the nature of astronomical observations - finite speed of light, which allows the observation of objects at large distance; technological advancement in the measurement instruments, such as the CCDs (Charged Coupled Devices) and the increasing performance of computational capabilities have contributed to expand the frontiers of Science in Astronomy. Discoveries such as the black holes and the dark matter were possible due to this. Following this progress, we could adventure that new important discoveries will occur in the coming years. This science contributes to a better understanding of the Universe and its history (Cosmology).

The observational nature of Science, especially relevant in domains such as Astronomy and Astrophysics sets important challenges in the data analysis domain; currently, limitations are perceived not only in the area of the uncertainties inherent in the existing state-of-the-art technologies associated with the instruments, but also in the amount of data retrieved and the need to obtain reliable quality flags to handle these associated uncertainties in an optimum manner. With the advances in computing frameworks, on-line catalogues are becoming interconnected and search motors are improving their capabilities. This e-Science platform is a good basis for the exploitation of AI methodologies to extract scientifically meaningful knowledge from on-line data. These aspects need to be carefully assessed when modelling and analysing observed data.

In this sense, the set of parameters which defines astrophysics problems normally contains high dimension and notable heterogeneity. This parameter set constitutes the basis of the information used in our understanding of the physics of the Universe. In addition to this, Astronomy and Astrophysics use, similarly as in any discipline, a specialized language. Knowledge engineering needs to take on board this specialized language in order to create an effective ontological platforms, as it is the case with IVOA (International Virtual Observatory Alliance), for example. All these aspects make astrophysics problems suitable candidates for the application of AI in a knowledge engineering frame.

Following the logic stated in [Schneider, 2014], we can contribute to the understanding of the Universe by understanding its elements such as galaxies. In this research, we contribute to this increasing understanding of the Universe by expanding knowledge acquisition capabilities.We use the astronomical data analysis and implement AI methodologies, mainly related to Data Mining and artificial vision, both used within a coordinated framework of associated knowledge engineering.

Researching in Artificial Intelligence (AI) usually involves other disciplines. Astronomy is one of

those disciplines in which the implementation of AI has become very relevant in recent years. One reason for this can be found in the increasing complexity in terms of models and quantity of acquired data which this area of Science has experienced over the last few years. In this research we worked with multi-band astronomical images, all large enough areas of deep cosmological fields which contains galaxies and stars. The information about the position (hereafter named astrometric) and the brightness (hereafter named photometric) of the objects found in these images is described in astronomical catalogues, which were also part of our input data set. One element frequently used in the classification of astronomical objects and especially in galaxies is their spectrum. Basically, the spectrum of a galaxy is a vector of information about how much energy is emitted by the galaxy in a specific wavelength. When we consider detections of objects in several wavelengths we talk about spectrum energy distribution (SED).
Traditionally, the classification and cataloguing of astronomical objects requires some heavy manual steps based on the astronomer's visual perception and classification criteria in order to identify the same object in the different bands, to determine the most probable separation of overlapped objects and to compute the source brightness, etc., all this with the added difficulty of data uncertainties. The considerable increase in data rate in recent years is challenging this traditional approach against a new one based strongly on a pipeline architecture with intensive computational functionalities, all this integrated in a knowledge engineering framework with the overall goal of automatically extracting knowledge from data. In this sense, we work with catalogues already existing in the astronomical community. The quality of the data of these catalogues is expressed through quality indicators which provide useful information in terms of the validity of the extracted data.
Therefore, the context of this research involves the processing of large astronomical data in order to achieve an effective automatic classification system of the data cubes along with their associated quality flags.
Two main hypotheses have driven the considerations of the approaches in this research and the decisions taken about the methods implemented:

- The quality of the data of astronomical catalogues is expressed by quality indicators. These indicators provide useful information in terms of validity of the measurements for various contexts of their usage.
- Currently limitations are perceived not only in the area of uncertainties derived from the technological boundaries of the measurement instruments, but also in the amount of data obtained.

We address in this research a multiple problem, which can be formulated as tasks derived from each specific objective, whose implementations are based on specific methods:

1. **Objective 1**: to improve the quality and effectiveness of the identification of the same astronomical source - galaxy in this case, across multiple wavelength images exposures with crowded populations.

   (a) **Task 1**: to automatically identify the cross-matched galaxies from the astrometric and photometric data of multi-wavelength catalogues.

   (b) **Method 1**: Bayesian model decision framework, where the model has been obtained with a simple convolution of the Spectral Energy Distribution (SED) templates with the appropriate filter along a wide range of redshifts.

2. **Objective 2**: to improve the quality, effectiveness and accuracy of the surface brightness computation of extended astronomical sources - galaxies in this case, from any kind of astronomical FITS image.

(a) **Task 2**: to determine the contour and compute the surface brightness of galaxies in the images of astronomical surveys.

(b) **Method 2**: two methods were used and benchmarked - (1) active contour without edges refined by iteratively adjusting (based on the outcomes of the previous iteration) the parameters involved in the differential equation of this method, and (2) MATLAB contouring algorithm with configurable brightness levels implementation and pixel exploration algorithm to determine the optimum contour.

Objective 3: to create an effective automatic classification system of labelling isolated versus non-isolated astronomical sources of diverse catalogues and FITS images.

(a) **Task 3**: to automatically discriminate isolated from non-isolated galaxies in the images of astronomical surveys.

(b) **Method 3**: to implement the Voronoi tessellation and a simple rule-based system based on the inclusion or not of the contour in its voronoi cell.

Therefore, the main focus of this multiple problem was first to optimize the relevant quality flags of any astronomical catalogue. The main outcomes of these three lines of research can be ingested into an ontology to refine the classification of the galaxies and by such to improve the quality of the knowledge acquired, which is the ultimate goal of the overall research presented here.
In order to effectively use the information, it is important to develop systems capable of archiving and accessing data efficiently. Nowadays, data archives constitute one of the corner stones in Astrophysics research. Intensive computation on the archived data takes place in parallel among various research communities.
Today it is a reality that before any large-scale astronomical survey programme starts producing data, a typical project life cycle for data processing and analysis is put in place. Therefore, concept, requirements, design and validation phases are sequentially taking place to develop a full data processing pipeline with the ultimate goal of producing scientifically meaningful information. In recent years we have observed more effective interaction between the disciplines of pure Astronomy, Astrophysics and that of Artificial Intelligence. This multidisciplinary synergy proves to be very effective in the context of large complex data processing and analysis.
Therefore the goal of our research has been to improve the cataloguing of galaxies in the areas of identification of multi-band cross-matches (considering best fit of spectral energy distribution (SED)), to improve the process of compute galaxies surface brightness for isolated galaxies and to identify overlapped and isolated objects, which altogether contribute to the refinement of the quality flags normally produced in every catalogue.
Thus, the implementation of this goal constitutes in itself an expert data analysis pipeline which is extensively described in Chapter 3 and validated in Chapter 4; the corresponding compilation of results for one area of the sky is detailed in Chapter 5. Finally, Chapter 6 presents the main conclusions and future lines of work to be derived from this research.
We can easily extend the foundations of the contributions presented in this research and detailed in the following chapters, to similar scientific scenarios in Astronomy or in other disciplines.
For the sake of external readers who are not necessarily expert in all these disciplines, dedicated bibliographic references for further readings relevant to this research are included as appropriate in this document.

Chapter 2

Motivation and State-of-the-Art

2.1 Introduction

This chapter introduces first an overall view of this research in terms of motivation, context and main objectives. The state-of-the-art connected to each objective was addressed in dedicated sections of this chapter. Each of these sections contains a bibliographic review of the current approaches and a brief justified discussion of the methodology chosen. In order to provide an efficient description of these objectives, we presented in a concise manner the relevant context and then the specific problem and its resolution is further elaborated in Chapter 3 through the description of the methodology implemented.
Today astronomical projects normally involve a wide spectrum of wavelengths and they may cover various astronomical surveys. In this framework the usual approach is to design, develop, implement and validate what is named a processing expert system pipeline which fulfils a set of scientific end user defined requirements.
The main objectives of this research is to improve the classification system of multi-band astronomical sources-galaxies. This is achieved working in four areas of improvement:

- To discriminate between isolated and overlapped sources for each band.
- To cross-match the sources from various bands.
- To refine the computation of brightness of the source.
- To integrate the above areas of improvements into a knowledge engineering framework.

All this together constitutes a pipeline for the automatic processing of massive data, in which each main step requires a thorough study and implementation of specific techniques. The justified choice of each technique used is presented in the following sections of this chapter.
Therefore, the ultimate goal of this research is to automatically extract useful information for the scientific astronomical community.
The main discipline considered in this research is Artificial Intelligence (AI), used in conjunction with Data Mining-Bayesian inference, computer vision-active contour, Data Mining-rule-based classification and knowledge engineering-ontology. The context in which AI has been applied is astrophysics-cosmology. Therefore, this research brings together two domains of knowledge (Astrophysics and

Artificial Intelligence) integrated in a joint framework capable of achieving goals only possible when these two disciplines are working together.
The remainder of this chapter encompasses a first section introducing the knowledge engineering discipline and within it the state of the art in terms of astronomy ontology in the context of the Virtual Observatory; then a section presenting the Data Mining discipline and within it the state-of-the-art of rule based systems and of Bayesian inference for astronomical cross matching is included; it follows a dedicated state of the art section on computer vision and focused on the problem of segmentation and finally a section about the most relevant aspects of Big Data and associated scientific challenges concludes the chapter. The state-of-the-art considered in this research is mainly related to the context of application in Astrophysics, Cosmology and Astronomy in general.

2.2 Knowledge Engineering. Ontology in Astronomy

Review of the State-of-the-Art

We observed in recent years an increment and extensive use of ontologies across various disciplines and contexts. For example, ontologies have become very familiar with the World Wide Web. The WWW Consortium (W3C) is developing the Resource Description Framework (RDF), a language for encoding knowledge on Web pages to make it comprehensible to electronic agents searching for information. As introduced in [Cambrésy et al., 2010], the semantic web and ontologies are emerging technologies which enable advanced knowledge management and sharing. Nowadays many disciplines develop standardised ontologies which domain experts can use to share and annotate information in their fields. Their application to astronomy can provide a powerful tool for the exchange of information not only between astronomers, but also between machines of software components, and of allowing inference engines to perform reasoning on an astronomical knowledge base.
We found some bibliographic updated references of Knowledge Engineering (KE) in the open Cambridge Journal *The Knowledge Engineering Review* (KER). This review is committed to the development of the field of AI and the clarification and dissemination of its methods and concepts. KER publishes analyses of high-quality surveys providing balanced but critical presentations of the original ideas; technical tutorials, reviews on bibliographic references, giving the contents of current journals in theoretical and applied AI.
[Cambrésy et al., 2010] presents the status about describing the knowledge on astronomical object types. Specifically, this ontology of defined concepts is designed to enable advanced reasoning on astronomical object types. Through this work, [Cambrésy et al., 2010]are exploring the possibilities of defined concept ontologies in the field of astronomy. Ontologies are structures representing and formalizing knowledge. Their usage ranges from basic classification to advanced inference and reasoning.
The International Virtual Observatory Alliance (IVOA), a worldwide scientific organisation formed in June 2002 has as its primarily objective to enable adequate access to data gathered by astronomical observatories. An information system allowing such access is called a **Virtual Observatory** (VO). This organisation focusses an extraordinary effort on defining standards to ensure interoperability of the various virtual observatory projects already existing or in development.
The IVOA is now composed of 19 VO projects from Argentina, Armenia, Australia, Brazil, Canada, China, Chile, South Africa, Europe, France, Germany, Hungary, India, Italy, Japan, Korea, Russia, Spain, the Ukraine, the United Kingdom and the United States.
Various working group distribute the tasks of the IVOA, being the most relevant ones for our research the following:

- **Applications**: mainly concerned with the software tools which astronomers use to access VO data and services for carrying out astronomical research.

- **Data Access Layer**: in charge of defining VO standards for the remote access. Some examples of rules documented in this working group are Simple Image Access Protocol, SIAP6, Simple Spectra Access Protocol, SSAP7 and Simple Cone Search, SCS8.
- **Data Modelling**: the primary activity is to provide a framework for the description of metadata attached to observed or simulated data. The focus of this working group is on logical relationships between metadata and identification of an architecture for handling them, all this considering how an astronomer would handle this information. This group is closely linked wit the Data Access Layer one.
- **Grid and Web Services**: this group explores and define the use of grid technologies and web services in the context of VO.
- **Resource Registry**: The Resource Registry Working Group defines the structure and interface of what is named the IVOA Registry. With this registry, an astronomer can locate and use any resource archived in IVOA space. Interoperability functionality is also part of this working group.
- **Semantics**: this team creates new standards for the interoperability among VO systems. The meaning or the interpretation of words, sentences, or other language forms in the context of astronomy is the primary focus here. They also study and evaluate the relationship between words, symbols and concepts and their representation in an ontology. Central methodologies and tools involve the use of natural language in astronomy, including queries and translations.
- **VO Query Language**: This team defines a universal Query Language. Its usage is in the distributed access of the Virtual Observatory framework.
- **VOTable**: in charge of the VOTable format, an XML standard for the interchange of data represented as a set of tables. This format is closer to the FITS Binary Table format.
- **Theory Interest Group**: During the IVOA Executive meeting in January 2004 in Garching, Germany, the IVOA Theory Interest Group intended to:
  - Provide a forum for discussing theory specific issues in a VO context.
  - Contribute to other IVOA working groups to ensure the inclusion of theory specific requirements.
  - Incorporate standard approaches defined in these groups when designing and implementing services on theoretical archives.
  - Define regular services relevant for archives.
  - Promote the development of services for comparing theoretical results with observations and vice versa.
  - Define significant milestones and assign specific tasks to interested parties.

The IVOA proposed a standard format for vocabularies based on the W3C's Resource Description Framework (RDF) and Simple Knowledge Organization System (SKOS). In this manner, IVOA will allow the various group to create a maintain their vocabularies while keeping the access possibility by the rest of the astronomical community. A list of current IVOA Recommendations and other technical documents are at $http://www.ivoa.net/Documents/$
[Gray et al., 2008] establishes a set of conventions for the creation, publication, use and manipulation of astronomical vocabularies within the Virtual Observatory, based upon the W3C's SKOS standard. According to [Gray et al., 2008], a number of astronomical vocabularies have been created, with a variety of goals and intended uses. Some examples are:

- The *Second Reference Dictionary of the Nomenclature of Celestial Objects*, which contains 500 paper pages of astronomical nomenclature.
- For decades professional journals have used a set of reasonably compatible keywords to help classifying the content of whole articles. These keywords have been analysed by Preite Martinez and Lesteven, who derived a set of common keywords constituting one of the potential bases for a further VO vocabulary. The same authors also attempted to derive a set of common concepts by analysing the contents of abstracts in journal articles, which should comprise a list of more up-to-date tokens/concepts than the old list of journal keywords. A similar but less formal attempt was made by Hessman for the VOEvent working group, resulting in a similar list.
- Astronomical databases generally use simple sets of keywords - sometimes hierarchically organized - to help users make queries. Two examples from very different contexts are the list of object types used in the Simbad database and the search keywords used in the educational Hands-On Universe image database portal.
- The Astronomical Outreach Imagery (AOI) Working Group has created a simple taxonomy for helping to classify images used for educational purposes.
- In 1993, Shobrook published an astronomy thesaurus endorsed by the IAU, [Shobbrook and Shobbrook, 1993]. This collection of nearly 3000 terms, in five languages, is a valuable resource, but has seen little use in recent years. Its very size, which gives it expressive power, is a disadvantage to the extent that it is consequently hard to use.
- The VO's Unified Content Descriptors (UCD) constitute the main controlled vocabulary of the IVOA and contain some taxonomic information. However, UCD has some features which supports its goals, but which make it difficult to use beyond the present applications of labelling VOTables: firstly, there is no standard means of identifying and processing the contents of the text-based reference document; secondly, the content cannot be openly extended beyond that set by a formal IVOA committee without going through the laborious and time-consuming negotiation process of extending the primary vocabulary itself; and thirdly, the UCD vocabulary is primarily concerned with data types and their processing, and only peripherally with astronomical objects (for example, it defines formal labels for RA, flux, and bandpass, but does not mention the Sun).

As concluded in [Gray et al., 2008], the goal of this standard is to show how an interoperable and computer-manipulable format can express vocabularies without difficulties. The machine reasoning capability is not a goal of this specification. Thus, the use case here is about searching for data, rather than representing the data itself, and for this, the looser semantics of a thesaurus are more appropriate than the formal ("is-a") definition of an ontology. More elaborate artefacts would be precious but require much more expensive work.

Choice of the Methodology

Knowledge Engineering (KE) was defined in 1983 by [Feigenbaum and McCorduck, 1983] as follows: "KE is an engineering discipline which involves integrating knowledge into computer systems to solve complex problems normally requiring a high level of human expertise". Also, as introduced at the International Conference on Knowledge Engineering and Ontology Development (KEOD) 2014, and defined in `http://www.definitions.net/definition/Knowledge%20engineering`, "Knowledge Engineering refers to all technical, scientific and social aspects involved in building, maintaining and using knowledge-based systems. KE is a multidisciplinary field,

bringing in concepts and methods from several computer science domains such as artificial intelligence, databases, expert systems, decision support systems, etc."
Ontology Development (OD) aims at building reusable semantic structures that can be informal vocabularies, catalogues, glossaries as well as more complex finite formal structures representing the entities within a domain and their relationships. Ontologies have been gaining interest and acceptance by computational audiences: formal ontologies are a form of software; thus software development methodologies can be adapted to serve ontology development.
Ontology-based systems are not very dependent on the evolution of the ontology. This means that when the astronomical knowledge evolves, one just has to update the ontology accordingly and the system exploiting it will take the updates into account. **Protégé** is a free open-source platform which provides the user community with a suite of tools for constructing domain models and knowledge-based applications with ontologies, such as the ones being created at IVOA.
An ontology uses a common vocabulary for researchers who need to share information in a domain. This vocabulary involves machine-interpretable definitions of concepts in the domain and relationships among them. The primary motivations for developing an ontology are the following:

- To share the common understanding of the structure of information among people or software agents. This is one of the most common goals in developing ontologies.
- To enable reuse of domain knowledge. This need of reuse was one of the driving forces behind the recent surge of the ontology. The modular feature of an ontology allows for an integration of various ontologies into a large one. Each of those integrating ontologies describes a portion of the whole domain under consideration.
- To make domain assumptions explicit. This explicit capability makes it possible to change these assumptions quickly if our knowledge about the field changes. Hard-coding assumptions make them not only hard to find and understand but also hard to modify. Also, exact specifications of domain knowledge are useful for new users who must learn the meaning of the terms used in the field.
- To separate domain knowledge from operational knowledge. This is another common use of ontologies. We can describe the task of configuring a product from its components according to a required specification and implement a programme that creates this configuration independently of the products and components themselves.
- To analyse domain knowledge. we need for that the availability of a declarative specification of the terms. Formal analysis of concepts is precious when both were attempting to reuse existing ontologies and extending them.

Often, an ontology of the domain is not a goal in itself. Developing an ontology involves defining a set of data and their structure for other programmes to use. Software agents, systems for problem-solving methods, etc. use ontologies and knowledge bases built from ontologies as data.
In practical terms, an ontology describes formally and explicitly concepts in a domain of knowledge, also named classes. This description also involves the properties of each concept describing features and attributes of the concept, and restrictions. A knowledge base is thus composed of an ontology, together with a set of individual instances of classes. Classes are the focus of most ontologies. They describe the concepts in the domain. A class can have subclasses which represent concepts that are mores specific. Slots represent properties of the classes and instances.
Therefore, in practical terms, developing an ontology includes:

- defining classes in the ontology,
- structuring the classes in a taxonomic hierarchy of subclass-superclass,
- defining slots and describing allowed values for these slots,

It is a fact acknowledged by IVOA and in general by the astronomy and astrophysics community that astrophysical concepts and quantities use a wide variety of names, identifications, classifications and associations, most of which cannot be described or labelled using IVOA Unified Content Descriptors (UCDs). Vocabularies are primarily associated with searching and browsing tasks. In ontologies, the domain is formally captured in a set of logical classes, typically related in a subclass hierarchy.
In terms of formal definitions, and as captured in [Gray et al., 2008], the standard ISO 5964: 1985 (or BS 6723:1985) defines a **controlled vocabulary** as "a prescribed list of headings each one having an assigned meaning (noting that controlled vocabularies are designed for use in classifying or indexing documents and for searching them" and a **thesaurus** as "a controlled vocabulary in which concepts are represented by preferred terms, formally organized so that paradigmatic relationships between the concepts are made explicit, and the preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms".
As indicated in [Gray et al., 2008], an astronomical ontology is necessary if we need a computer "understand" something of the domain. There has been some progress towards creating an ontology of astronomical object types to meet this need. However, there are distinct use cases for letting human users find resources of interest through search and navigation of the information space. The most appropriate mechanism to meet these use cases derives from what is named *controlled vocabularies, taxonomies and thesauri*.
As explained in [Gray et al., 2008], there are multiple vocabularies in use, describing a broad range of resources of interest to professional and amateur astronomers, and to members of the public. These various dictionaries use different terms and different relationships.
One approach to this problem is to achieve a consensus in the vocabulary which draws terms from the various existing vocabularies to create a new one which can express anything its users might desire. The problem with this is that such an effort would be very expensive, both in terms of time and effort. This effort would be not only on the part of creating and maintaining it but also to the potential users, who have to learn using the new terms correctly.
The alternative approach to the problem is that, rather than deprecating the existence of multiple overlapping vocabularies, IVOA should embrace it, help interest groups formalise as many of vocabularies as are appropriate, and standardize the process of formally declaring the relationships between them. This means that:

- The various vocabularies are allowed to evolve separately, on their own time-scales, managed either by the IVOA individual Working Groups within the IVOA, or by third parties;
- specialised vocabularies can be developed and maintained by the community with the most knowledge about a specific topic, thus ensuring that the vocabulary will have the most appropriate breadth, depth, and precision;
- Users can choose the vocabulary or combination of vocabularies most appropriate to their situation, either when annotating resources, or when querying them; and
- We can retain the previous investments made in vocabularies by users and resource owners.

The IVOA community has reached a consensus that formalised vocabularies should be published at least in SKOS (Simple Knowledge Organization System) format, a W3C draft standard application of RDF to the field of knowledge organization. SKOS draws on long experience within the library and information science communities, to address a well-defined set of problems to do with the indexing and retrieval of information and resources.
It is important to note that, as explained in [Gray et al., 2008], a vocabulary is not an ontology. It has lighter and looser semantics than an ontology, and is specialised for the restricted case of resource retrieval. It is easy to transfer the vocabulary relationship information from SKOS to a formal ontological format such as Ontology Web Language (OWL).

2.3 Knowledge Extraction

In the last decades, with the evolution of information technologies, such as cloud computing and data storage, along with the increase in storage capabilities, the information is mainly reachable in electronic support, shared through the internet by a diverse set of users communities. As referred in [Ball and Brunner, 2010], the increase in the amount of available data is creating a digital world, in which extracting meaningful information is a challenge itself.
Data Mining is a relatively recent discipline which includes various data analysis techniques. The ability to extract meaningful information from measurements is very useful (if not a necessity in order) to gain a better understanding of Nature. In general terms, as indicated in [Orallo et al., 2004], this information will reduce the uncertainties about a specific aspect of the reality and therefore this will effectively support the extraction of knowledge and the decision making task. In Data Mining the understanding of the data and of the model is key to implementing effective techniques in the process of extracting, analyzing and providing interpretation to the data. Therefore as described in [Orallo et al., 2004] there are two main targets within the Data Mining discipline: one is to reach an efficient capability of working with big volume of raw data and the other target is to use adequate techniques to analyse and classify the data and to extract useful and new knowledge.
Some of the limitations indicated in [Ball and Brunner, 2010] associated with the use of a Data Mining approach in large astronomical data sets are the following:

- In the astronomical context it is usually inappropriate to extrapolate beyond a specific parameter space.
- The error associated with astronomical data measurements is not always explicitly taken into account in many Data Mining algorithms.
- The optimal configuration of the parameters space is usually not intuitive.
- Many Data Mining algorithms scale, for n objects, as n^2, or even worse, in both memory and CPU time.
- Data mining is in itself an entire field of study with strong connections to statistics and computing. The full exploitation of this discipline for Science return requires a substantial learning curve.
- Many astronomical data sets are larger than can be held in a single machine memory and this is becoming more and more frequent nowadays. Therefore the link between Data Mining and parallel computing is necessary in order to succeed in the handling of astronomical large data sets and/or intensive computation.

Astronomy is one of the disciplines in which the increase of measurement capabilities and consequent data acquisition has demanded the clear need of Data Mining. This need has been motivated as enumerated in [Ball and Brunner, 2010] by aspects such as the automation of extracting scientific information from very large data sets; the simplicity of some Data Mining algorithms capable of producing quick useful results; the possibility to incorporate prior information and the algorithmic capability of highlighting patterns in a data set that otherwise might not be noticed by a human analysis.
Regarding Data Mining tools in the field of Astronomy, there is a wide variety of them. One example is the Sky Image cataloguing and Analysis System, SKICAT, for catalogue production and analysis of catalogues from digitized sky surveys. For a more detailed reading on this, [Ball and Brunner, 2010] recommends [Kamath, 2009].
As pointed out in [Ball and Brunner, 2010], and linked to the object detection process, visualisation of data is an important part of the scientific process and the current combined framework of Tera/Petascale and Data Mining suggests the need for a synergy between different disciplines. In

this sense, collaboration between computer science and other disciplines has resulted in progress in various areas of scientific visualisation. It is likely that such collaborations will continue to increase in importance.
In general a collaboration between the scientist and the expert in Data Mining can produce effective results in the implementation of Data Mining in disciplines, such as Astronomy. Nowadays multi-disciplinary workshops are being organized all over the World and the new technologies provide very good support in these interactions (i.e, web conferences, etc.). One example of this kind of forums was the 2012/2013 SAMSI programme on Statistical and Computational Methodology for Massive Data-sets where experts from different disciplines - Astronomy, Meteorology, Medicine, Data mining, Parallel Computing, Artificial Intelligence etc - collaborated in a joint effort framework to bring together sound solutions in the extraction of useful information from data to serve to the progress of Science.

2.3.1 Labelling of Astronomical Sources

Review of the State-of-the-Art

A Voronoi diagram performs a decomposition of a given space based on the distances to a specified family of objects (subsets) in that space. We name these objects generators, and with each object one associates a corresponding Voronoi cell, namely the set of all points in the given space whose distance to the given object is not larger than their distance to any other objects. It is named after Georgy Voronoi, and is also called a Voronoi tessellation, a Voronoi decomposition, or a Dirichlet tessellation (after Lejeune Dirichlet).
Properties

- Delaunay triangulation is the dual graph for a Voronoi diagram (in the case of a Euclidean space with point sites) for the same set of points.
- The closest pair of points corresponds to two adjacent cells in the Voronoi diagram.
- Assuming that the plane is Euclidean and given a set of different points in it, we say that two points are adjacent on the convex hull if and only if their Voronoi cells share an infinitely long side.
- Under relatively general conditions (i.e., the space being infinite dimensional and uniformly convex) Voronoi cells enjoy stability property; for example, a change caused by some geometrical modification (translation, distortion), produces small variations in the shape of the Voronoi cells in the neighbourhood. This property is the geometric stability of Voronoi diagrams. This feature does not hold in general (e.g. if space is non-Euclidean)

As introduced in [Okabe, 1992], it is quite possible that the concept of the Voronoi diagram has been used since antiquity. For example, in his treatment of cosmic fragmentation in both *Le Monde de Mr. Descartes, ou Le Traité de la Lumière* published in 1644 and in part III of *Principia Philosophie* also published in 1644, Descartes uses Voronoi-like diagrams to show the disposition of matter in the solar system and its environments, and it is possible that such figures were not uncommon at that time. The first time - that the authors of [Okabe, 1992] are aware of - that the concept of Voronoi tessellation is introduced appears in the work of Dirichlet (1850) and Voronoi (1908), who, in their studies on positive definite quadratic forms, considered a special form of the Voronoi diagram. Dirichlet treated two and three dimensional cases, whereas Voronoi examined the general n-dimensional case. As a result of this, the name Voronoi diagram started to be used over the years in a wide spectrum of different science and social disciplines.
As described in [Okabe, 1992], by early 1970 some algorithms had been developed to construct

Voronoi diagrams in two and three dimensions. The paper of Shamos and Hoey (1975) presented an algorithm for constructing the Voronoi diagram and introduced aspects such as the identification of the nearest neighbour to each point.
Hereafter, Voronoi tessellations are being widely adopted in many scientific and engineering fields to describe spatial distribution. Astronomy is not an exception. This technique is very important to the spatial database domain since it is a fundamental concept in partitioning a large spatial data set.
In order to provide the reader with a flavour of the wide spectrum of applications of the Voronoi tessellation, some examples of implementation of Voronoi tessellations in astronomy are presented below.
Reviewing the history a little, as indicated in [van de Weygaert, 1994], Kiang seems to be the first one to apply Voronoi tessellations to astrophysics as documented in [Kiang, 1966]. He used the Voronoi tessellations for his study of the mass spectrum obtained in the fragmentation of interstellar clouds. Matsuda and Shima were the first ones to propose the use of Voronoi tessellations in cosmology, as described in [Matsuda and Shima, 1984]. They pointed out the similarity between two-dimensional Voronoi tessellations and the outcome of numerical clustering simulations in a neutrino-dominated Universe published in [Melott et al., 1983]. Icke and Van de Weygaert (IW) (1987) introduced the Voronoi tessellations into cosmology. In their paper, [Icke and van de Weygaert, 1987], they studied the statistical properties of two-dimensional Voronoi tessellations. After finishing a three-dimensional geometrical Voronoi algorithm, IW found that Voronoi tessellations also have some unusual features of clustering. Yoshioka and Ikeuchi (1989) confirmed this in [Yoshioka and Ikeuchi, 1989].
According to [van de Weygaert, 1994], the main contribution of the Voronoi tessellation in a cosmological context is that it has a very powerful mathematical model which contains unusual properties of clustering in a cellular structure. Moreover, the Voronoi tessellation may yield a reasonable accurate quantitative description of some specific physical models of structure formation. An alternative use of Voronoi tessellations is as statistical descriptors of galactic distribution. Finally, [van de Weygaert, 1994] concludes that considering that the galaxy distribution on large scales resembles a cellular structure, with voids playing a prominent role, models based on a hierarchical build-up of structure in the Universe seem to be particularly successful in explaining the structure from a scale of galaxies up to a range of cluster and groups of galaxies. The beauty of the Voronoi tessellation is that, although its definition is simple, it leads to a structure of great richness which as yet is still largely unexplored.
As indicated in [van de Weygaert, 1994], in its cosmological context the Voronoi tessellation can be considered as the skeleton of a galaxy distribution. Such a skeleton would result from an idealized model of structure formation in the Universe dominated by the expansion of under-dense regions into the matter distribution. In this way a better understanding of the cellular clustering system can be obtained and this will be very useful when interpreting more complicated and realistic models of clustering.
The concept of identifying a footprint in the sky which contains a set of detections from different surveys has been elaborated and implemented in [Budavári et al., 2007] through a web tool [1]. This tool can complement the pipeline presented in this research by allowing a clear visualisation of the area of the sky where the pipeline is going to be executed. Due to the computational cost of the pipeline, the input image at each execution is normally a tile of a mosaic image and, thus being much smaller than the full sky, the planar Voronoi can be executed in each tile image. The footprint tool described in [Budavári et al., 2007] determines mainly the coverage of the sky for specific measurements, without entering in a detailed analysis of the potential isolated sources within the footprint itself.

[1] http://www.voservices.net/footprint

A variant of the Voronoi diagrams called curved Voronoi diagrams is described in [Ramella et al., 2001] and [Boissonnat et al., 2006]. [Boissonnat et al., 2006] describes in detail two general mechanisms to obtain effective algorithms for some classes of curved Voronoi diagrams.

Choice of the Methodology

One of the areas studied in the Artificial Intelligence discipline is the problem of classification. As indicated in [Bertin, 2001], classification is a very important initial step in the scientific process, providing a mechanism for organizing the information in an effective way to apply the scientific method.
As indicated in [White, 1997], any classification method uses a set of features or parameters to characterise each object. There are two main approaches in classification: supervised and unsupervised classification. The very first step of classification in Astronomy can be at the object detection step, which is normally done at some signal to noise threshold. Various Data Mining algorithms, embedded in specific software package, such as SExtractor, are used for that purpose.
As explained in [Bertin and Arnouts, 1996] Source-Extractor (SExtractor) is a program that builds a catalogue of objects from an astronomical image. This tool is particularly oriented towards reduction of large galaxy-survey data, but it also performs well on moderately crowded star fields. A more detailed description of this software package can be found in [Bertin and Arnouts, 1996].
We use the Voronoi diagram as a preliminary tessellation of galaxies detected in a portion of the sky, where in a first approximation the celestial coordinates of the galaxies catalogue will play the role of generators.
As part of this research, we will implement the simplest tessellation case, which in [Okabe, 1992] is called a planar ordinary Voronoi diagram and Chapter 3 defines it formally. In general terms, given a finite set of points $\{p_1, ..., p_n\}$ in the Euclidean plane, we are interested in finding for each point p_k its corresponding Voronoi cell $V(p_k)$. The points of this Voronoi cell will have the distance to p_k no greater than their distance to any other site. We obtain that cell from the intersection of half-spaces and is hence a convex polygon. The segments of the Voronoi diagram are all the points in the plane that are equidistant to the two nearest sites. Voronoi vertices are points equidistant to three (or more) sites.
Our images are pre-processed by the catalogue tool SExtractor and in the case of the images used here we can consider that after pre-processing the astrometric information (celestial coordinates) is kept consistent among images obtained from diverse telescopes and/or instruments. On this basis, we also assume that the level of astrometric uncertainty will be adequate to allow the effective implementation of our processing pipeline. With this in mind, we chose the first-order Voronoi tessellation as the most appropriate method to be implemented as a first step towards identifying isolated sources from their celestial positions in the fits images. Regarding computational cost aspects, an incremental algorithm can be performed on tiled images by using the Voronoi tessellation approach. The boundary problem is a well-known natural complication; one possible solution is to mirror the points near the boundary of the overall image; another possible solution is to identify those points on the boundary and to consider them in the following tiled image where they are not on the boundary. The points in a boundary of a 2D Voronoi tessellation have at least one vertex of its Voronoi cell in the infinity.
As introduced above, one of the main objectives of this research consists of identifying the isolated sources from the astrometric information offered by a 2D astronomical image. This image will in this case be mainly composed of galaxies in a specific band and it is accompanied by a catalogue obtained from a pre-processing step run by SExtractor tool. In this sense, our knowledge of the astronomical catalogues and images is very rich and therefore neither supervised nor unsupervised techniques are the best option here. Instead, an initial tessellation and a simple rule-based system can cope with the problem of labelling areas of the sky where a specific condition is true. Our input

information is the catalogue data obtained from the celestial coordinates of each galaxy of the image under consideration. The Voronoi tessellation seems to be an adequate approach to preliminary splitting the image in areas based on the object astrometric positions. It retains the simplicity of the ordinary planar Voronoi diagram and at the same time the efficient capability of processing any kind of galaxies spatial distributions allowing a clear assessment of the neighbourhood of each galaxy.
Considering the objectives of this research and the type of images used, the curved Voronoi diagram is not a good choice here because it would be based on the region of our sources (which normally have blurred boundaries and very commonly are overlapped with others); instead, the traditional Voronoi tessellation is based on the coordinates of the astrometric centre of each source, which offer a much more reliable basis for the tessellation.

2.3.2 Cross-Matching of Astronomical Sources

Review of the State-of-the-Art

As explained in [Wall and Jenkins, 2012] decision is a fundamental part of Science, and we implement this process generally through comparison. We then need to accompany each measure by an error estimate. This error estimate is a measure of range (expressed regarding probability) that encompasses the belief of the actual value of the parameter.
On the other hand, [Wall and Jenkins, 2012] explains that attempting to understand astrophysics and cosmology requires some reconsideration of the classical scientific method because this approach based on repetition of experimentally reproduced results does not seem to apply here. The inability to repeat experiments or to reproduce the same space-time conditions has caused some of the greatest errors of inference in the field of astronomy.
In [Wall and Jenkins, 2012] the following stages are identified in making Science:

- Observations: process for obtaining and recording data
- Reduction: cleaning of the data by removing undesirable effects of the measurement system, and calibration of the data.
- Analysis: obtaining numbers which represent the cleaned and calibrated data allowing comparison or modelling.
- Conclusion: implementation of a process to reach a decision.
- Reflection: identification of what has been learned; analysis of the decision reached in terms of plausibility.

The concept of probability is crucial in the decision process; the distribution of a specific parameter normalized to have an area equal of 1.0 is the measured probability density function, often called the probability distribution. The tails of such distributions contain little area and it is these tails that are often key for the decisions.
Statistical inference methods represent the bridge between the model and the data. Common statistical problems in astronomy fall into categories of tasks such as data compression, classification, feature extraction, parameter estimation, model selection, etc.
The astrostatistics community has excellent bibliographic references these days with useful reviews of Bayesian inference methods in cosmology.
There is an ongoing debate between the advocates of a “frequentist” approach and the Bayesian methodology. The frequentist approach interprets probability as the frequency of the outcome of a repeatable experiment, whereas the Bayesian methodology interprets the probability more generally and includes a degree of belief. In the Bayesian approach the choice of priors may strongly affect the inference. However it is an “honest” approach in the sense that all the assumptions are explicitly

spelled out in a logical manner.
The characterisation of systematics error constitutes an important part of research in astronomy. The contribution of the different types of systematic errors in the error budget should be incorporated into the statistical analysis accordingly.
As explained in [Wall and Jenkins, 2012], once we make the measurements, it is usual to try to correlate the observations with other results. Various reasons can be found for this: (1) to check that other observers' measurements are reasonably in line with ours, (2) to check that our measurements are reasonable themselves and (3) to test the validity of a hypothesis, perhaps one for which we planned the observations. It is important to remember here that correlation does not prove a causal connection. From this point, we identify two different approaches for proceeding: (1) Bayesian and (2) non-Bayesian.
The Bayesian approach uses Bayes' theorem to extract the probability distribution for the correlation coefficient from the likelihood of the data and suitable priors. The alternative approach starts by regarding the correlation coefficient as a fixed quantity, not a variable about which probabilistic statements might be made.
When performing data sample comparison, we are carrying out hypothesis testing. We can divide classical methods of hypothesis testing into two categories: (1) parametric and (2) non-parametric. Bayesian methods require the notion of a known distribution.
As explained in [Gregory, 2010], in Science, available information is always incomplete and therefore knowledge of Nature is necessarily probabilistic. Two different approaches exist: in traditional statistics, we identify the probability of an event with the relative frequency of its occurrence. We name this as the "frequentist" view. However, a new perception of this probability has appeared in recent years. With this new notion, the mathematical rules of probability are valid logic principles for conducting inference about any proposition or hypothesis of interest. This viewpoint is called the Bayesian Probability Theory (BPT) and is playing an increasingly important role in physics and astronomy. Therefore, the Bayesian inference encodes, with a probability distribution, our uncertainty about some model parameter or set of competing theories, based on the current state of information.
Bayesian inference is about inferring the laws of Nature from experimental data and prior information. Bayesian analysis can improve the estimates of the model parameters by encompassing both inductive and deductive logic. The source of our general understanding comes from the definition of theoretical models capable of explaining what we observe as well as making testable predictions. [Gregory, 2010] focuses on what happens at the interface between the predictions of scientific models and the data from the latest experiments. The data are always limited in accuracy and completeness; therefore we are unable to employ deductive reasoning to prove or disapprove the theory. In this sense, Bayesian inference provides a means for assessing the plausibility of one or more competing models, and for estimating the model parameters and their uncertainties, the latter being carried out as part of what is called data analysis.
The following provides a review of the current state-of-the-art in the resolution of the cross-matching problem.
[Budavári and Szalay, 2008] present a general probabilistic formalism for cross-identifying astronomical point sources in multiple observations. In this sense, a Bayesian framework is defined for the object matching, where consideration can be given not only to spatial information but also to physical properties, such as colours, redshifts and luminosity. In [Budavári and Szalay, 2008] a description of a proposed recursive algorithm is provided in order to evaluate the Bayes factor over a set of catalogues with circular uncertainties in positions. This work also implements the Bayesian framework defined in terms of spatial information (astrometry). [Roseboom et al., 2009] present an implementation of a combined Bayes Factor for an astrometric and photometric cross-matching of catalogues as described in [Budavári and Szalay, 2008]. In this thesis a simple prior is used and before using the approach presented in [Roseboom et al., 2009] the Spectral Energy Distribution (SED)

must be carefully determined, tested and optimised on simulated or well-known real datasets. A Spectral Energy Distribution (SED) is a graph of the energy emitted by an object (flux) as function of different wavelengths. The purpose of this graphic is to see how much energy is produced by the object as a function of frequency or wavelength. To read more about this, please refer to [Schneider, 2014], [Romanishin, 2006] and [Budding and Demircan, 2007].
Following the line of thought provided in [Budavári and Szalay, 2008] and [Roseboom et al., 2009], an implementation of a refined framework for a photometric cross-matching and the associated SED fitting methodology, the analysis of uncertainties and the impact of different priors in the outcome is provided in Chapter 3.
[Heinis et al., 2009] present the results from the investigation on the quality of associations of astronomical sources from multi-wavelength observations using realistic simulated detections. They describe a general method to build mock-up catalogues for studying associations and comparing statistics of cross-identification.
An extension of the work presented in [Budavári and Szalay, 2008] for cross-matching is proposed in [Kerekes et al., 2010] for the case of dynamic sources; i.e., sources which move in the sky. From the results [Kerekes et al., 2010] conclude that the use of more realistic prior distributions within the Bayesian framework can improve cross-identification of moving astronomical objects.
An example of implementing the cross-matching Bayesian framework proposed in [Budavári and Szalay, 2008] for a real data set is described in [Lubow et al., 2011], presenting the result of cross-identifying sources observed in different Hubble Space Telescope (HST) exposures (see [Jäger, 1990]). As explained in [Lubow et al., 2011], for each set of overlapping exposures, a trimmed source list of possible cross-matches is determined using a Bayesian method. Then a novel algorithm is applied for shifting the exposures into common alignment. This is iteratively repeated until no further improvement is obtained. With this approach, results show that the author(s) have been able to obtain mean relative astrometric shifts between exposures of order of a few milliseconds.
More recently, photometric classification techniques for the star-galaxy classification in multi-band optical imaging have been studied in [Fadely et al., 2012]; this includes supervised data-driven Support Vector Machines (SVM) and unsupervised spectral energy distribution fitting (i.e. Maximum Likelihood (ML) and Hierarchical Bayesian (HB) which learns the prior distribution of template probabilities from the data). These methods are implemented and the results are benchmarked in order to find the algorithm which yields the best performance. Results in [Fadely et al., 2012] conclude that SVM requires training data to classify unknown sources; ML and HB don't. Simulated and realistic scenarios have been considered.
The importance of catalogue cross-matching, widely used in astronomy, is stressed in [Pineau et al., 2011] by acknowledging the need for efficient, reliable and scalable cross-catalogue matching with forthcoming projects which will produce large catalogues where astronomers are expected to dig for rare objects and perform statistical analysis and classification. The formalism proposed in [Pineau et al., 2011], which support more than simple nearest-neighbour search, addresses this challenge. They resolve the scalability problem by partitioning the sky using the Hierarchical Equal Area iso-Latitude Pixelation (HEALPix) scheme and independently processing each cell. This pixelation produces a subdivision of a spherical surface in which each pixel covers the same surface area as every other pixel. More details can be found in [Gorski et al., 2005].The whole process can, therefore, run on a single computer, but could also use clusters of machines to cross-match future vast surveys in a reasonable time.
One example which demonstrates the feasibility and utility of large database cross-correlation in discovering rare interesting objects can be found in [Metchev et al., 2008] where it is reported that new L and T drafts have been found in a cross-match of the SDSS data release 1, see [Abazajian et al., 2003] and 2MASS, see [Skrutskie et al., 2006]. As explained in [Metchev et al., 2008], the simultaneous search of the two databases effectively makes it possible to explore the combined databases to a greater completeness level. As result in this case, two new T dwarfs are found, in

addition to the 13 already known and a peculiar potentially metal-poor L2 dwarf with unusually blue near-IR colors IS also found. These discoveries highlight the utility of simultaneous database cross-correlation in searching for rare objects. [Metchev et al., 2008] states that a cross-correlation of the SDSS and 2MASS databases may allow us to probe not only deeper, but also cooler discoveries than in either survey alone. The ability to cross-correlate large astronomical databases is one of the main technological goals of the National Virtual Observatory (NVO). Despite the great scrutiny with which the area had already been explored for T dwarfs, both of the new T dwarfs had previously been overlooked. One possible cause is that suspect photometry flags might have raised in SDSS. [Metchev et al., 2008] concludes that the discovery of the two new T dwarfs demonstrates the superior sensitivity to ultra cool dwarfs that can be attained by simultaneously cross-correlating large optical and near-IR databases.

An example of Bayesian inference of parameters can be found in [Bailer-Jones, 2011] where a Bayesian method is introduced for estimating the intrinsic parameters of a star and its line-of-sight extinction. It uses both photometry and parallaxes in a self-consistent manner in order to provide a non-parametric posterior probability distribution over the parameters. The author of [Bailer-Jones, 2011] demonstrates this method by using it to estimate effective temperature and extinction from BVJHK data for a set of artificially reddened Hipparcos stars. [Bailer-Jones, 2011] states from the results that this method can easily be extended to incorporate the estimation of other parameters, in particular metallicity and surface gravity, making it particularly suitable for the analysis of the 10^9 stars from Gaia. Finally [Bailer-Jones, 2011] concludes that this method has significant advantages over traditional pattern recognition methods, such as neural networks or support vector machines. As explained in [Bailer-Jones, 2011], the method developed in this paper does not need training data, or modelling of the parallax or apparent magnitude. Pattern recognition methods normally also give just a single solution and are incapable of naturally providing probability distributions over parameters. They likewise cannot incorporate prior information. These methods cannot explicitly take into account the constraints from physical background information, with the consequent risk of yielding non-physical solutions. The only advantage of the traditional methods is their speed but this is hardly an advantage compared to the usefulness of the outcomes.

In [Chapin et al., 2004] another example of Bayesian parametric inference is presented for photometric redshift of (sub)milimiter selected galaxies. In this work a luminosity function is used as prior.

[Krughoff et al., 2005] presents a case study in which the popular source extraction code "SExtractor" has been deployed in a web service, which in addition provides cross-matching capabilities between the resultant catalogues and other previously-published data sets. This prototype has been developed through the SkyPortal web service provided by OpenSkyQuery.

As indicated in [Krughoff et al., 2005], this proposal constitutes an important step in the incremental development of web services. XML and specifically SOAP are chosen as platform and language independent development environment. [Krughoff et al., 2005] enumerates as follows some of the intermediate uses of this service: (1) photometric calibration by comparison with catalogues with known photometry; (2) identification of optical counterparts to sources in other bands; (3) search for high redshift objects by looking at photometric drop-outs; (4) star-galaxy separation; (5) variability studies, etc.

In [Malik et al., 2003] a Web prototype - named SkyQuery- for practical implementation of queries in federated databases is presented. As indicated in [Malik et al., 2003], a federated database is a collection of cooperating but autonomous component archives behaving like a single integrated database. SkyQuery supports a complex query called cross match query which is a spatial query in charge of finding the matched objects across archives if they correspond to the same astronomical body. Finally [Malik et al., 2003] enumerates a few technical difficulties in the implementation of the prototype and the way around these.

Following this, [Nieto-Santisteban et al., 2007] analyses the issues and requirements that an environment needs to consider when aiming to enable large-scale astronomical science. The context of

this was the primary mission of the National Virtual Observatory (NVO). It describes a workbench environment - CasJobs- where astronomers can cross-correlate, compare and analyse vast amounts of data sharing their results with the community. The basis of the solution proposed in [Nieto-Santisteban et al., 2007] is the commercial Relational Database Management Systems, and SQL expresses the analysis tasks. [Nieto-Santisteban et al., 2007] also proposes an indexing algorithm, named Zone, explaining its fundamental role in enabling parallelism and cross-matching of vast data sets.
The bibliographic references where bayesian inference is used in Astronomy has experienced an important ramping up. For example, Gaia survey has started using it, as described in [Gaia Collaboration et al., 2017]. Some basic online tools exist for the implementation of cross-matching following a proximity position criteria[1]; the virtual observatory has also released a dedicated cross-matching tool[2] which allows one to cross-identify sources coming from two different tables. In this case, cross-match is performed on the basis of the position or else taking into account the error ellipses of the sources. In the area of Galaxy Photometry, the tool ARCHANGEL is described in [Schombert, 2011].

Choice of the Methodology

There are two main reasons why the concept of probability is important in astronomy:

- Astronomical measurements, as any measurement, are subject to both systematic and random errors. We have to express these errors as precisely and usefully as possible.
- The inability to carry out experiments on our subject matter leads us to draw conclusions by contrasting properties of controlled samples.

With the exponential growth of data in astronomy, there is a great need for further interaction between astronomers and experts in other fields. With the recent multi-communities interaction in scientific forums, astro-statistics has become a “respectable” discipline in its own right. In this sense, Bayesian approaches are more commonly used and in better co-existence with frequentist methods. Today there is more awareness of model selection methods and therefore computer intensive methods are more popular.
As described in [Ball and Brunner, 2010], the combination of available information with incomplete coverage of the possible phenomena suggests that a probabilistic approach, either involving priors or semi-supervised, will generally be the most suitable approach, because it will allow the use of existing information, while interpreting a new phenomena objectively.
As pointed out by [Wall and Jenkins, 2012], the main reasons why statistical inference based on known probability distributions does not work are the following:

- We are observing and measuring as experiments are being run out there in the Universe, not by us. The underlying distributions may be far from known or understood. The exception to this might be non-parametric statistics, methods that do not require knowledge of the underlying distributions.
- We may have to deal with a small number of samples.
- The use of scales other than numerical requires in most cases the use of non-parametric methods.

As indicated in [Loredo, 1990], the Bayesian approach to probability theory compared to the frequentist one, is more closely related to our intuitively reasoning in front of uncertainties. [Loredo, 1990] enumerates and briefly describes the main pitfalls of the frequentist approach:

[1] http://www.ict.csiro.au/staff/Robert.Power/projects/CM/ps/cm.htm
[2] http://www.usvao.org/2011/11/21/cds-cross-match-service-release/

- Arbitrariness and Subjectivity: assessing hypotheses is one of the principles of probability theory. However, the Frequentist theory had to develop alternative ways of accomplishing it without calculating the probabilities of these hypotheses. The solution to this problem triggered the creation of the statistics discipline.
- Comparison with Intuition: according to the Bayes' theorem, the posterior probability allows the assessment of a variety of hypotheses; this posterior probability depends on the prior probability of the hypothesis and the probability of the one data set observed. In contrast, the frequentist reasoning reversed the roles of hypothesis and data. However, due to the absence of the concept of the probability of a hypothesis, the frequentist approach needs to assume that a single one of the space of hypotheses is the truth and find ways to assess this decision
- Randomness vs. Uncertainty: frequentist theory is forced to base inferences on possible data because data, and not hypotheses are considered to be "random variables". A close inspection of the notion of randomness reveals further difficulties with the frequentist viewpoint.
- The frequentist Failure: in the past, significant deficiencies in the Bayesian theory motivated surface of the frequentist approach to probability theory. Unfortunately, frequentist theory addressed these deficiencies only by burying them under a superficially more objective facade. The apparent arbitrariness of the probability axioms motivated the approach to the frequentist theory. The notion of frequency to define the probability intended to solve that problem. But this definition forbade the use of Bayes' Theorem for the analysis of hypotheses, and as a result, this frequency theory cannot yield unique solutions to well-defined problems. The addition of an arbitrary and various set of principles and criteria did not improve the issue.
  The subjective nature of Bayesian probability assignment was an important motivation for the frequentist theory. However, the notion of randomness widely used in the frequentist approach is itself subjective, dependent on one's state of knowledge in a manner very similar to that of Bayesian probability.
  Finally, frequentist theory is badly at odds with the manner in which we intuitively reason in the presence of uncertainty. Rather than evaluate a variety of hypotheses in the light of the available evidence, the theory attempts to evaluate a single hypothesis by considering a range of hypothetical data. It also ignores any prior information one may have regarding the possible hypotheses.
  In conclusion, frequentist theory has failed to address the problems that motivated it. Though its successful use for a broad set of problems is indisputable, it can give contradictory results for some problems. This situation indicates the need for a better theory.

[Gregory, 2010] enumerates as the main advantages of the Bayesian approach as follows:

- Provides an elegantly simple and rational approach to answering, in an optimal way, any scientific question for a given state of information. This method is in contrast to the cookbook approach to traditional statistical analysis. For some problems, a Bayesian analysis may simply lead to a simple statistic.
- It calculates the probability of hypothesis directly.
- It incorporates relevant prior information through Bayes' theorem. This feature is one of the greatest strengths of Bayesian analysis. For data with high signal-to-noise ratio, the use of relevant prior information in a Bayesian analysis can result in an improvement of various orders of magnitude in the model parameter estimation.
- Provides a way of eliminating nuisance parameters through marginalisation. For some problems, the marginalisation can be performed analytically, permitting certain calculations to become computationally tractable.

- It retains a significant, powerful way of assessing competing theories at the forefront of Science by automatically quantifying Occam's razor. Occam's razor is a principle attributed to the medieval philosopher William of Occam. The principle states that one should not make more assumptions than the minimum needed. This principle guides us towards the selection of the simplest model to explain a given phenomenon.

- Provides a way of incorporating the effects of systematic errors arising from both the measurement operation and theoretical model predictions.

As indicated in [Hobson, 2010], in Science, perhaps especially in branches such as cosmology, we create models designed to make sense of the data we have collected. The mathematics discipline formalises these models; but even here, we do not find the absolute truth. Therefore, as detailed in [Hobson, 2010], our question in Science is not "Is this hypothetical model correct?", But " Is this model better than the alternative?". Rational assessment of different models is the central subject of Bayesian methods

According to [Gregory, 2010] the Bayesian Probability Theory (BPT) is playing an increasingly important role in physics and astronomy. As indicated in [Trotta, 2008] Bayesian methods have proven to be vastly superior to more traditional statistical approaches, yielding higher efficiency and a consistent approach to the problem of uncertainty. The increase in computing capabilities through parallel processing, Graphical Processing Units (GPUs), cloud computing and other emerging technologies has started to support in a suitable manner the use of mathematical frameworks, unfeasible a few years ago, such as Bayesian inference and its derived intense numerical computation for complex problems. According to [Trotta, 2008], cosmology is perhaps among the latest disciplines to have embraced Bayesian methods, a development mainly driven by the data explosion of the last decade, as demonstrated by the dramatic increase in the number of cosmology and astrophysics papers using Bayesian Theory. (see [Trotta, 2008]).
Astronomical cross-identification of the same source in multi-wavelength scenarios has been a very well-known problem for years. The deployment of an efficient framework where automatic classification pipelines can run is an area currently being explored in fields such as astrostatistics and Data Mining. As described in [Marquez et al., 2014], the problem of identifying the same astronomical source in different exposures and across instruments has long been studied for its statistical and computational complexity. Any astronomical data set generally contains a large number of attributes for each object. The algorithms deal with each of these attributes as dimensions of the mathematical problem; this is why it is important to implement a size reduction in such a way that there is a right balance between the minimum number of attributes and the maximum quantity of information.
We typically consider the coordinates of the sources in the sky to decide whether they belong to the same object, but crowded areas often yield degenerate cases when multiple possible matching configurations have similar likelihoods. Accordingly the consideration of additional information, such as photometry seems to be a good approach here.
Probabilistic cross-matching of surveys has been implemented by Budavari (see [Budavári and Szalay, 2008]). [Ball and Brunner, 2010] states that full PDFs have been shown to improve performance in the photometric detection of galaxy clusters, due to the increased signal-to-noise ratio. Proper characterization of errors will be of great importance for future surveys as the probabilistic approach becomes more important.
We have chosen here the use of Bayesian inference applied to multi-wavelength photometric measurements in order to obtain a more accurate framework when dealing with cross-matching of sources among various instruments of diverse wavelengths.

2.4 Computer Vision

In the following paragraphs, we offer a brief introduction to the context and particularities of astronomical images and their associated imaging methodologies. Light can be measured and defined by its wavelength. However, colours and their perception by humans are subjective because of their interpretation by the eye-brain-mind. The "accurate" recording of colour in images is complicated because our perception is also very adaptive to the intensity and colour of the light under which we observe the subject.
Images of astronomical objects are usually taken with electronic detectors such as a Charge Coupled Devices (CCD). Typically the astronomical images are measured through specific filters. Different detectors and telescopes usually have different sensitivities to different colours (wavelengths). Filters can either be broad-band or narrow-band. For example, a broad-band filter lets a broad range of colours pass through the instrument. Conversely, a narrow-band filter typically only lets a small wavelength span through it, restricting the radiation that passes through the device and by that, allowing the astronomers to investigate specific atomic processes in the object.
Image processing is by its nature a very computationally demanding task; this is explained in [Howell, 2006]. As a result, efficiency is a major concern when designing image-processing algorithms.
Noise is one of the main limitations when addressing effectiveness in image and signal processing. Therefore, reduction of noise is one important part of image processing discipline. A detailed description of the source of noise relevant to the astronomical images which we use is shown in [Marquez and Sarro, 2013].
As introduced in [Howell, 2006], CCDs are the state-of-the-art detectors in many fields of observational science, such as astronomy and cosmology in particular. The professional astronomer is nowadays using telescopes and evaluating data in measured in all known wavelengths. Imaging, Astrometry, Photometry and Spectroscopy are the main methods when using CCDs instruments. Imaging, Astrometry, Photometry and Spectroscopy are the most relevant methods when using the CCD device in Astronomy.
Typically the information collected by a CCD device is displayed via a computer. Each pixel in the CCD acts as an electrically isolated portion of the silicon array and is capable of incoming photon collection, photo-electrons storage, and CCD readout transmitted to an associated computer as a digital number. As detailed in [Solomon and Breckon, 2011], the word pixel is an abbreviation of a picture element and refers to the smallest part of a digital image. It contains a numerical value which is the core unit of information at a given spatial resolution and quantization level. Any camera is of fixed spatial resolutions (number of pixels making up the image) and has a fixed depth per pixel (number of bits used to represent each pixel). The fundamental concept in digitization is that of quantization; this means the mapping of the continuous signal from the scene to a discrete number of spatially-organized points (pixels), each with a finite representational capacity (pixel depth). In general, as long as sufficient samples are taken, a spatially-quantized image achieves non-noticeable errors in the picture.
In general, a telescope represents the image of an object field of view (FOV) as a weighted sum of 2D impulse functions and then expresses the image plane field as the weighted sum over these impulse functions. This is known, as named in [Howell, 2006] as the superposition principle, which is valid for linear systems as the one used astronomical observation. The images of the individual object-plane impulse functions are called Point Spread Functions (PSFs), and it reflects the notion that a mathematical point of light in the object plane is physically spread out forming a finite shape in the image plane. The PSF describes the way that information on the object function is spread as a result of recording the data and is a deterministic function that operate in the presence of noise. The fact that the PSF is typically determined entirely by the imaging system allows us to describe the image by knowing the optical properties of the measurement system. This process is usually formulated by a convolution equation. In astronomy, knowing the PSF of the measuring device is

very important for restoring the original image using deconvolution.
Therefore, as indicated in [Solomon and Breckon, 2011], the image formation process can be summarised as a small number of key elements. In general, a digital image S can be formalised as a mathematical model comprising a functional representation of the scene (the object function o) and that of the capture process (the (PSF) p). Additionally, the image will contain additive noise n. These are essentially combined as follows to form an image: $s = p * o + n$ where s is the image, p is the PSF and n the noise; $*$ represents the convolution operator.
In many contexts, we can consider the PSF as the extended blob in an image that represents an unresolved object. The degree of blurring of the point object is a common measure of the quality of an imaging system. In non-coherent imaging systems, such as telescopes, the image formation process is linear in power and described by linear system theory. Thus the imaging of an object A is unaffected by the imaging of another object B and vice versa. We can consider the image of a complex object as a convolution of it and the PSF.
The object function describes the object which is being imaged and the way in which the light is reflected from it back into the imaging instrument.
Noise is a non-deterministic function which we can only describe with some statistical distribution.
A good or sharp imaging system will generally possess a narrow PSF, whereas a poor imaging system will have a broad PSF, which has the effect of considerably overlapping the output responses of neighbouring points in the input.

2.4.1 Segmentation in Astronomical Images. Active contour

Figures 2.1 and 2.2 show a graphical illustration of the scale of observations that we obtain when modern instruments explore and produce images of astronomical deep fielda. It corresponds to the so-called eXtreme Deep Field (XDF): the combination of 10 years of NASA Hubble Space Telescope, from a patch of sky centred on the Hubble Ultra Deep Field. The XDF is a small fraction of the angular diameter of the full moon.

Review of the State-of-the-Art

[Solomon and Breckon, 2011] define image segmentation as a generic process by which an image is subdivided into its mutually exclusive constituent regions or objects. Completely autonomous segmentation is one of the most challenging tasks in computer vision and it remains an active field of image processing and machine vision research.
The segmentation process in astronomical images becomes a necessary step in a classification pipeline, also regarding quality flags. Usually we call foreground to the segmented objects and background to the rest of the image. The fundamental question regarding the problem of segmentation as proposed in [Solomon and Breckon, 2011] is: "What relationship must a given pixel have with its neighbours and other pixels in the image to decide if it belongs to one region or another?". This central question is usually approached in image segmentation through one of the two following ideas:

- Edge/boundary methods. The basis of this approach is the detection of edges as a means of identifying the boundary between regions. It looks for sharp differences between a group of pixels.

- Region-based methods. This approach assigns pixels to a given area based on their degree of mutual similarity.

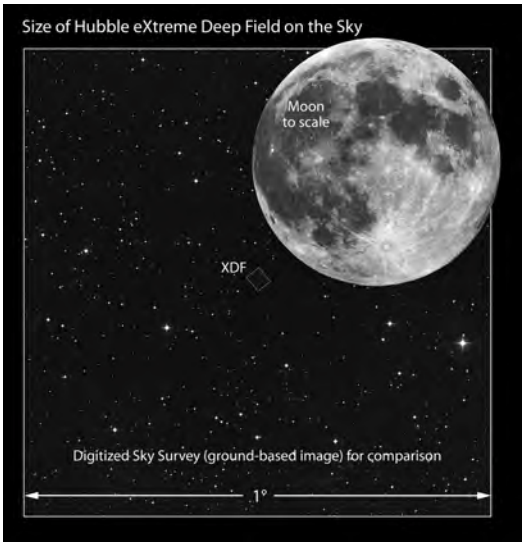

Figure 2.1: This illustration compares the angular size of the XDF field to the angular size of the full moon. A finger held at an arm's length would appear to be about twice the width of the moon in this image. Note that this illustration does not show the actual observation of the XDF relative to the location of the moon. (Illustration Credit: NASA; ESA; and Z. Levay, STScI; Moon Image Credit: T. Rector; I. Dell'Antonio/NOAO/AURA/NSF.)

Figure 2.2: The XDF is the deepest image of the sky ever obtained and reveals the faintest and most distant galaxies ever seen. Credit: NASA; ESA; G. Illingworth, D. Magee, and P. Oesch, University of California, Santa Cruz; R. Bouwens, Leiden University; and the HUDF09 Team.

Several methods have been developed for image segmentation. Information about relevant properties of the image, such as colour, texture, motion, .etc, will determine which method is the most appropriate for each image. We can identify the following group of methods in the image segmentation problem:

- Intensity threshold: the basic idea, as described in [Solomon and Breckon, 2011], is very simple. We choose some threshold value such that pixels above this value belong to one region and those pixels below the threshold belong to another region. The outcome of this is a binary image.
- Histogram-based methods: as indicated in [Solomon and Breckon, 2011], a histogram is computed from all the pixels in the image and the peaks and the valleys are used to identify clusters. Colour or intensity can be used for measurement purposes.
- Region growing and region splitting: in this method, described in [Solomon and Breckon, 2011], pixels are grouped into larger regions based on their similarity according to predefined similarity criteria. In order to effectively define the similarity criteria, it is necessary to consider the spatial adjacency between pixels. Region splitting essentially employs a similar philosophy, but is the reverse approach to region growing. In this case at the starting point the whole image is considered as a single region which is then successively broken into smaller regions until any further subdivision results in regions falling below some chosen threshold.
- Split-and-merge algorithm: The algorithm, described in [Solomon and Breckon, 2011], divides into two successive stages. The aim is to break the image into a set of disjoint regions each of which is regular within itself. First the whole image is considered as the initial area of interest; then it is decided whether all pixels belonging to the region satisfy some similarity criterion. If so, the area of interest corresponds to a region in the image and is labelled; if not, then the area of interest is split into four equal sub-areas, each of them considered as areas of interest in turn. This process ends when no further splitting occurs.
- Edge detection: this is one of the most important and widely-studied aspects of image processing. As indicated in [Solomon and Breckon, 2011], if we can find the boundary of an object by locating all its edges, then we have effectively segmented it. Edges are simply regions of intensity transition between one object and another. However, despite its conceptual simplicity, edge detection remains an active field of research. The use of gradient differential filters constitutes the basis of the most edge detectors.
- Partial differential equation-based methods: curve propagation is a popular technique here with many different applications. Typically, as explained in [Solomon and Breckon, 2011], an initial curve evolves towards the lowest potential of a cost function. The minimisation of the cost function is not an obvious problem, and it imposes certain constraints typically expressed in geometrical terms of the evolving curve. The following are the most relevant methods used with this approach to segmentation:
  - Parametric methods: Lagrangian techniques are based on parametrising the contour according to some sampling strategy and then evolving each element according to image and internal terms. The original "purely parametric" formulation is attributable to Kass and Terzopoulos, and known as "snakes", published in [Kass et al., 1988]. Such methods are fast and efficient, however, presents drawbacks: limitations in the choice of sampling strategy, the intrinsic geometric properties of the curve, topology changes (curve splitting and merging), etc. Nowadays, efficient "discretized" formulations have been developed to address these limitations while maintaining high efficiency. In both cases, energy minimisation is generally conducted using a steepest-gradient descent, whereby derivatives are computed using, e.g., finite differences.

  - Level set method: This method was initially proposed by Osher and Sethian in [Osher and Sethian, 1988] to track moving interfaces and spread across various imaging domains in the late nineties. It resolves the implicit propagation technique efficiently in the problem of the curve, surface, etc. A signed function represents the evolving contour, where its zero level is the actual contour. Then, according to the motion equation of the contour, one can easily derive a similar flow for the implicit surface which, when applied to the zero-level, will reflect the propagation of the contour. There are significant and numerous advantages in the level set method: it is implicit, it is parameter-free, it provides a direct way to estimate the geometric properties of the evolving structure, it can change topology, and it is intrinsic. Furthermore, it can be used to define an optimization framework, as Zhao, Merriman and Osher proposed in [Zhao et al., 1996]. An obvious conclusion from this is that this method is a very convenient one for addressing numerous applications of computer vision. Furthermore, efficient implementations are achieved using this method with a different level set data structures.
  - Fast marching method: The fast marching method, described in [Sethian, 1996], has been used in image segmentation, and it has achieved a considerable improvement by permitting a both positive and negative propagation speed in an approach called the generalised fast marching method.

- Graph partitioning methods: These methods can effectively be used for image segmentation. Here, as described in [Wu and Leahy, 1993], a graph models the image. This graph has no weight and no direction. This method associates a pixel or a group of pixels with nodes and edges which define the (dis)similarity between the adjacent pixels. Then a partition takes place according to a criterion designed to model "good" clusters. Each partition of the nodes (pixels) output from these algorithms is considered an object segment in the image.

- Watershed segmentation: this approach, presented in [Beucher and Lantuejoul, 1979] attempts to separate touching objects, which is one of the most complicated image processing operations. The central idea of the Watershed transformation, as described in [Beucher and Meyer, 1992], is to consider that the gradient magnitude of an image is a topographic surface. Therefore, pixels with the highest gradient magnitude intensities (GMIs) (watershed lines), represent the region boundaries. Considering this analogy, the pixels enclosed by the same watershed line will flow naturally downhill to a Common Local Intensity Minimum (LIM). A segment will be composed of the pixels draining to a common minimum form a catch basin.

- Multi-scale segmentation: Image segmentation, as described [Gauch, 1999], is computed at multiple scales in space and sometimes propagated from coarse to fine scales. The criteria for segmentation is typically arbitrarily complex and may take into account global as well as local criteria. An important notion here is that each region must be connected in some sense.

In what follows we briefly describe implementations of the above methodologies for specific real cases in the astronomy field.
One of the simplest approaches, as described in [Cristo et al., 2008], is based on establishing an image threshold where all pixels above certain intensity level are considered objects of the image and those pixels below the threshold belong to the background. The typical astronomical images contain diffuse boundaries of the sources, and the intensity level of object and background is not always the same for the different areas of the image. Due to this, an active contour methodology is frequently used because this technique allows a preliminary contour to evolve towards an optimum partition of the image, this means that region-based segmentation energy can be defined in a local manner. In [Verma et al., 2011] an edge based active contour model using adaptive threshold and ant colony optimisation is presented. The target is to obtain a well-constructed edge map of the image. The end points obtained using adaptive thresholds are calculated, and the ants are placed at

these points. The local variation in pixel intensity guides the movement of the ants. The probability factor of only undetected neighbouring pixels is taken into consideration while moving an ant to the next probable edge pixel.
In [Osher and Fedkiw, 2006] a brief summary of traditional segmentation methods is provided. As indicated, the basic idea in active contour models (or snakes), initially proposed by Kass et al in [Kass et al., 1988], is to evolve curves according to constraints from a given image, to detect objects in that image. These curves are affected by internal and external forces which are defined so that the snake will conform to the boundary of the image under study. The active contour models are based on the theory of surface evolution and geometric flows, as indicated in [Pan et al., 2011]. A level set formulation was proposed in Caselles et al. in [Caselles et al., 1997]. This formulation allows topological changes and geometrical flexibility and has been quite successful in two and three spatial dimensions. Then, in a sequence of papers beginning with [Chan and Vese, 1999], the authors propose different variations of the active contour model without Edge Stopping Function (ESF). The stopping term is based on the Mumford-Shah segmentation technique introduced in [Mumford and Shah, 1989]. Another feature of these proposed models is that the interior contours are automatically detected, and the initial curve can be located anywhere in the image.
The work described in [Chan and Vese, 1999] introduces the Mumford-Shah functional for segmentation and the level sets new model for active contours to detect objects. This is based on curve evolution techniques. This model is a combination of classical active contour models or snakes, using mean curvature motion methods, and the Mumford-Shah model for segmentation. The objective is then to minimise an energy which can be seen as a particular case of the minimal partition problem. This model is capable of detecting objects whose boundaries are not necessarily defined by gradients. As explained in [Chan and Vese, 1999], in all the classical active contour models, the stop condition for the evolving curve is an edge detector. In practice, the contour may pass through the border of the object due to the discretization of the gradient used in the edge function. In addition to this, if the image is too noisy, then the isotropic smoothing Gaussian has to be strong, which will smoothen the edges, too. [Chan and Vese, 1999] proposes a new active contour model, without stopping edge-function, which is, therefore, capable of detecting contours both with or without gradient. Later on, [Pan et al., 2011] presents an improvement related to increasing robustness in the location of the initial contour. In addition to that, a Gaussian Regularizing Level Set Method (GRLSM) is used to reduce the computational cost. This improvement removes the need for reinitialization.
The problem of overlapped objects in image segmentation has been known for a long time, and its solution is not evident. Active contours are limited in their ability to segment overlapping objects. [Fatakdawala et al., 2010] Presents a new segmentation method, Expectation-Maximization (EM) driven Geodesic Active Contour with Overlap Resolution (EMaGACOR), which is applied to initialize a Geodesic Active Contour (GAC) automatically. This method includes a novel technique for splitting contours via identification of high concavity points for resolving overlapping structures using the definition of edge path which defines the paths through relevant edge points within the contour while ensuring the optimal split. Thus EMaGACOR is presented as an efficient, robust, reproducible and accurate segmentation technique. The GAC model was introduced by Caselles et al. (cited in [Fatakdawala et al., 2010]), and it allowed for computation of minimal distance curves. Considerable advances in active contour models are taking place. Probabilistic models have recently been employed to drive segmentation techniques. The assessment of the performance of this EMaGACOR method is based on a set of metrics which evaluate features such as Sensitivity to Noise (SN), Positive Predictive Value (PPV) and Overlap Detection Ratio (OR).
A scheme for the detection of object boundaries is presented in [Caselles et al., 1997]. The technique is based on active contours which evolve in time according to intrinsic geometric measures of the image. This approach is based on the relation between active contours and the geodesics or minimal distance curves. The implementation of this geodesic approach to object segmentation allows one to establish a connection between the classical snakes based on energy minimization and the geometric

active contours based on curve evolution theory. As detailed in [Caselles et al., 1997], the model here is given by a geometric flow (PDE), based on mean curvature motion. This model is based on a curve evolution approach and not an energy minimization one.

Choice of the Methodology

A typical astronomical image contains a nearly flat background with objects that can be point-source or extended ones. Noise in these images is always present due to many factors (more detailed information can be found in [Schneider, 2014], [Romanishin, 2006] and [Budding and Demircan, 2007]).
Astronomical sources detected with an instrument and represented in a digital image usually present blurred contours where the separation between foreground (the source identified) and the background is not evident. The images considered in this thesis are obtained from different telescope instruments and at different wavelengths, and the intensity level will not be homogeneous even within a specific image. In addition to that, the boundary of the foreground objects will be diffuse and the contour detection becomes challenging in many cases. The uncertainties play also an important role because they have a heavy impact on photon counting and therefore on the foreground object brightness computation.
Traditionally, manual operations have been taking place where the astronomer uses various contour algorithms in a non-automatic case-by-case mode. The Kron ellipses is one typical traditional approach offered in SExtractor where a manual step to determine the aperture is necessary to compute realistic source brightness. As part of our research we intend to optimise the segmentation of the astronomical images automatically and in that way to refine the computation of surface brightness. The images that we will consider here correspond to deep cosmological fields where the segmentation of faint sources needs to be performed. The intensity level is heterogeneous across the image and between images; thus the intensity level of the background will differ from one area of the image to another and the noise level poses a real challenge for the faint objects. Finally, due to the nature of capturing the objects (by photon counting after calibration of the instrument) and the nature of the foreground object themselves (deep field galaxies, i.e., objects with high redshift at bands where the signal-to-noise ratio might not be very high), the frontier or contour of the object will appear diffuse or blurred. The images we will work with retain a level of complexity which require careful study in the selection of the best candidate for the segmentation methodology.
Taking into consideration the features of our astronomical deep field images, we have developed a segmentation methodology based on the region-based level set one. The reason for this choice is that we do not expect sharp difference in brightness for the boundaries of the objects in our images and, therefore, the region based methodology is more appropriate than the boundary methodology. Then, within the region-based approach, the level set method, as one of the partial differential equations methods, allows an important degree of flexibility in terms of contour evolution, which is very relevant for the inhomogeneous nature of our images. Thus, this method compared to the other ones retains important advantages compatible with the complexity of our images, while retains simplicity compared to other solutions, such as the Geodesic contour for example.

2.5 Big Data in Astronomy

As explained in [Scott et al., 2005], following Moore's law, in which computer performance has increased exponentially over the last several decades, the coming decade will probably face petascale computing. Much of the performance increase in the past decade has been driven by increases in processor (CPU) clock frequency, but this rate has now slowed due to physical limitations to the sizes of components, and more importantly to power consumption and energy dissipation. It

has therefore become more economical to manufacture chips with multiple processor cores.
As indicated in [Scott et al., 2005], it has been suggested that any algorithm which scales beyond $N \log N$ will rapidly be considered infeasible. The rapid evolution in microprocessor-distributed architecture, interconnection technology and software development is leading to a parallel processing approach to the simulation and solving of complex problems.
Many algorithms exist for the implementation of Data Mining on parallel computer systems but are not widely used within Science, as compared to the commercial sector.
One challenge of Data Mining is its introduction to parallel computing where the Data Mining algorithms normally needs to be substantially adapted to the new hardware at the level of pseudo code, requiring a consequent additional effort.
Historically, the primary objective of parallelism was to speed up computational processes regarding time and memory, characterised by **Amdahl's law**. This law defines the speedup of a task run at a fixed workload for a system whose resources are improved. It is as follows:

$$S_{latency}(s) = \frac{1}{(1-p) + \frac{p}{s}} \tag{2.1}$$

Where:

- $S_{latency}$ is the theoretical speed up in latency considering the whole task;
- s is the speed up in latency of the execution of the part of the task that benefits from the improvement in the system;
- p is the percentage of the execution time of the whole task.

Then **Gustafson-Barsis' law** focussed on scalability, i.e., keeping the time constant as the problem grows in size. This law gives the theoretical speedup in latency of the execution of a task at a fixed run time considering the improvement in the system. It follows::

$$S_{latency}(s) = 1 - p + sp \tag{2.2}$$

Where:

- $S_{latency}$ is the theoretical speed up in latency of the execution of the whole task;
- s is the speed up in latency of the execution of the part of the task that benefits from the improvement of the resources of the system;
- p is the percentage of the execution workload of the whole task concerning the part that benefits from the improvement of the resources of the system before the improvement

This led to incorporating more processors into the computational chain, starting then the cluster era. Recently, the Grid and Clouds ideas intend to incorporate the parallelism concept into a higher level. Furthermore, preliminary ideas about quantum computing, optical computing and biochip structures are starting to be proposed as the next step of the information technology evolution.
In general the performance of a computer code is the number of floating point operations that it an execute in a given time unit. Performance is often described in terms of millions of Floating Point Operations Per Second (mega FLOPS or MFLOPS) or billions (Giga FLOPS or GFLOPS). A new unit is defined, Cray, which equals 10^9 floating point operations per second (one floating point operation per nanosecond).
An example of petascale systems is the Blue Waters system at the National Centre for Supercomputing Applications (NCSA). This supercomputer provides a sustained performance of 1 petaflop, and

"it is composed of more than 237 Cray XE6 cabinets and 32 cabinets of the Cray XK7 supercomputer with NVIDIA Tesla Kepler GPU computing capability". https://www.infragard-illinois.org/)
As indicated above, time and memory are two key aspects to be considered in the context of parallel computing. Certainly one important feature of any algorithm is the work/memory ratio reflecting the relationship between the amount of work done (floating point operations) and the amount of memory that must be accessed (memory locations referenced, either reads or writes).
Parallel Algorithm Design Methods. In general, the design methodology for a parallel algorithm consists for the following four steps:

- *Partitioning*: the problem is broken down into fine-grain tasks, maximizing the number of tasks which can be executed simultaneously.
- *Communication analysis*: determines what communications are required among the tasks. An appropriate output from this step can take the form of a task graph with fine-grain tasks like nodes and communication channels as edges.
- *Granularity control (or agglomeration)*: aims to reduce communication requirements by combining groups of fine-grain tasks into fewer but larger coarse-grain tasks.
- *Mapping*: assigns coarse-grain tasks to processors, trying to achieve an optimal trade off between communication costs and degree of parallelism.

From MATLAB R2012 on, it includes a Parallel Computing Toolbox which allows the performance of parallel computations on multi core computers, GPUs, and computer clusters. The toolbox provides 12 workers - MATLAB computational engines - to execute applications locally on a multi core desktop.
In addition to this Parallel Computing Toolbox, MATLAB R2012 Distributed Computing Server Software allows the running of as many MATLAB workers on remote clusters of computers as licensing permits.
Many applications involve multiple segments of code, some of which are repetitive. Often "for-loops" are used to solve these cases. The ability to execute code in parallel can significantly improve performance in many cases. In this sense, the Parallel Computing Toolbox software enhances the performance of such loop executions by allowing several MATLAB workers to execute single loop iterations simultaneously. Even running all local workers on the same machine as the client, we can achieve a significant performance improvement with a multi-core architecture.
When working interactively in a MATLAB session, it is possible to offload work to a MATLAB worker session to run as a batch job. The command to perform this job is asynchronous, which means that the client MATLAB session is not blocked, and therefore the continuation of work in an interactive session is possible.
If an array is too large for the computer's memory, Parallel Computing Toolbox software allows that array to be distributed among multiple MATLAB workers, so that each worker contains only a part of the array.

Considerations in this Research

The astronomical images are usually very large. Nowadays astronomical surveys cover large areas of the sky and the acquisition data rate has entered the realm of Big Data. During the compilation of results from the pipeline proposed in this research we acknowledged certain difficulties related to both the large size of input data and the time of computation on a very small portion of the sky.
In terms of the large size of input data, the main difficulties were associated to the limitation in computing capabilities of the system used. One Matlab instance installed on a nominal configuration of dual CPU processor was not powerful enough to compute for example our complete pipeline on

one of the 113 tiles of the COSMOS image. Thus we implemented a Matlab script to select areas within a tile transferring the pixel coordinates of the full image into those reduced areas.
Another limitation was directly connected with the size of data obtained as outcome of the processing of the pipeline. The Matlab workspace and associated results were archived in separate redundant hard disks of the size of various Terabytes.
Regarding processing speed, the use of multiple indented loops due to a multiple dimension problem (for the case of photometric cross- matching algorithm) or to a recursive approach (for the case of the active contour algorithm) was the main cause of low speed in the processing algorithms. This behaviour was found in both the cross-matching and the active contour modules of the pipelines. Parallel processing would have been ineffective in the case of indented for loops, thus we only converted selected parts of our algorithm into the parallel process frame of Matlab.
Another difficulty encountered was the low speed in transferring archived data from external hard disks into the processing environment where the Matlab algorithms were running. This was particularly evident when it was required to retrieve some contextual results from past executions. However, promising results in both size and speed were achieved by splitting the area of processing into smaller ones. Thus, the solution developed in the context of this research is versatile enough to be adapted to the current and future set up in view of the Big Data paradigm.

Chapter 3

Expert System for Galaxies

3.1 Introduction

The traditional classification and cataloguing of astronomical objects requires some difficult manual steps based on the astronomer perception and classification criteria to determine the most probable separation of overlapped objects, to perform probabilistic cross-matching by identifying the same object in the different bands, etc. The considerable increase in data rate in the recent years is challenging this traditional manual method. The challenger is a new approach strongly based on data processing pipeline architectures involving intense and intelligent computational capabilities.

Survey campaigns nowadays handle observations for which the measurement instruments can be effectively exposed to the deep field area of observation for an extended period of time, yielding challenging images with crowded populations, overlapped objects, and faint detection, all this reaching the limits of instrument performance.

In this respect, the use of Artificial Intelligence (AI) techniques can optimise astronomical data processing and analysis by considering aspects related to the labelling of isolated versus contaminated astronomical objects, the identification of the same astronomical objects in different exposures of various instruments, and the measurement of the surface brightness of extended astronomical sources. This becomes especially relevant for those objects in which the measurements retain low-quality flags.

The analysis of multi-band images of cosmological fields presents several technical difficulties related to the assembly of consecutive knowledge extraction tasks from source identification to photometric measurements, cross-matching and source labelling. We present a new expert system to carry out some of these tasks in an automatic way using techniques borrowed from the field of AI.

We aim to provide the scientific community a tool to i) label sources in an astronomical image according to the probability that each one is blended or contaminated with surrounding sources; ii) extract photometric measurements of extended objects using contour-based approaches; and iii) carry out a probabilistic cross-match of sources from different images and bands, using all astrophotometric information available.

The new method proposed for the extraction of the photometric measurements is based on a well-known iterative scheme in the field of artificial vision for contouring the extended sources. The classification of sources into the various isolation categories is based on the Voronoi tessellation of

the images, and the relative position of the sources contours within each tessella. Finally, the probabilistic cross-match of sources from different images is based on a Bayesian inference process based on astro-photometric data.
This Section 3.1 addresses the methodologies used for the resolution of the problem detailed, the design of the new expert system and its implementation. Three main sections are part of this chapter, each one addressing each of the contributions of this research. A concise description of the complete pipeline is also provided here.
NOTE: from now on the word pipeline will be understood as synonymous of the word Galaxy Cataloguing Expert System (GCES).

3.2 Description of the Problem

The problem considered in this research contains various dimensions each of them pointing to a line of resolution where a specific AI methodology is used.
Therefore, this problem can be broken down into the following sub-problems, defined as the objectives of this research:

- To improve the quality and effectiveness of the identification of the same astronomical source, galaxies in this case, across multiple wavelength image exposures with crowded populations.
- To improve the effectiveness and accuracy of the surface brightness computation of extended astronomical sources, galaxies in this case, from any kind of astronomical FITS image.
- To create an effective automatic classification system of isolated versus non-isolated astronomical sources of diverse catalogues and FITS images.

The effective resolution of the above objectives is achieved through the implementation of an integrated but configurable automatic data processing pipeline which brings together disciplines such as astronomy and artificial intelligence (AI). This brings a three-axis impact and contribution; firstly in terms of the successful creation of a versatile software architecture which effectively takes benefits of a joint framework involving knowledge from astronomy and artificial intelligence (as a branch of computer science); secondly, through the public presentation of the contributions of individual blocks of this pipeline, considerable interest has clearly been perceived within the scientific community in the approaches proposed here and the results obtained; thirdly an incremental continuous improvement through iterative use of automatic processing on Big Data has been explored, achieving promising results. With all this, it was shown how the implementation of this multi-domain approach can improve the quality of astronomical data processing and analysis.
Overall, the pipeline described in this document is mainly intended to complement any existing nominal pipeline in the area of astronomical data processing and analysis.
The development environment chosen for the implementation of this pipeline was MATLAB, mainly for its extensive set of mathematical resources, such as libraries, and features, such as vectorisation and, in summary, its quick prototyping capabilities.
Many of the images used in astronomy are produced with a format (FITS) which comprises a considerable amount of information related to the image and the measurement setup. This raw data, also referred to in some data processing contexts as level 1 data, is the input used for the calibration of the images, e.g. the one described in [Capak et al., 2007], and it is strongly linked to the particularities of the instrument in charge of the measurements. The processing of this level 1 information in terms of reduction, rectification, calibration etc., yields what is named in some contexts level 2 data and this is the data used normally in the form of catalogue as input for astrophysical and cosmological consideration which normally yields scientifically usable products. The process involved in obtaining level 2 data from level 1 data is normally a complex one and the quality of the result is usually recorded in the form of quality flags included in the data set catalogue. Thus, the processing

and analysis assessment of level 2 data produces scientifically-meaningful information mainly for the astronomical science community. AI methodologies are being used in this research with both level 1 and level 2 data as part of the astronomy-computer science synergy framework which has been built up in recent years.
The following Figure 3.1 shows the block diagrams of the main generic steps involved in our astronomical data processing and analysis pipeline. Astronomical catalogues contain information about the quality of the level 2 data. This shows specific areas with low quality flags due to various reasons. The software presented here is based on dedicated studies already conducted in each of the areas which we are going to enumerate.
As introduced in [Marquez et al., 2014], most astronomical studies today rely on a fundamental processing step that creates the associations called cross-matching. The Big Data accompanying recent astronomical studies make it impossible to perform extensive manual steps in the cross-matching process. For cases where the uncertainties in the results are high, the combination of additional measurements, such as photometry, normally yield an improvement in the quality of the results.
The goal of [Marquez et al., 2014] was to design and implement an automatic process which uses measured fluxes to improve the quality of cross-matches. The methodology used was Bayesian hypotheses testing which provides a clear framework for combining different measurements in the decision-making process. In this research we extended the astrometric cross-matching concept implemented in [Budavári and Szalay, 2008] by using additional measurements (photometry).
Another area which clearly impacts the quality of the level 2 data is the computation of object surface brightness. Current astronomical surveys normally deal with multi-wavelength high-precision images, where the accuracy of the photometric measurements is quite good. The calibration process and its strong link with the instrument via the level 1 data is very relevant here. Noise is one of the main challenges to face when computing the surface brightness of extended sources such as galaxies. This is also identified in [Capak et al., 2007] and a separate dedicated study on using parametric Bayesian inference was also performed and documented in [Marquez and Sarro, 2013]. From the results with low quality masks we observed that the use of traditional tools for the surface brightness of galaxies do not always yield the best result. In this research we explore the use of a dynamic iterative contour identification methodology to automatically determine the optimum contour of extended astronomical sources and thus their surface brightness.
Typically, the astronomical multi-wavelength images contain numerous crowded areas where the confusion between objects is highly probable. This additional challenge in the area of data processing and analysis requires a robust step in classifying and labelling the objects as being candidates to isolated or non-isolated sources. In this sense, [Marquez, 2012] describes a rule-based system which implements this labelling. This system has been improved by using the active contour approach, as shown in the next sections of this chapter.
Following the high-level blocks of a generic pipeline, such as the one shown in 3.1, we define ingestion points where our optimisation pipeline can run. This is shown in Figure 3.2. Each of the orange blocks shown in 3.2 corresponds to a specific topic to which this research contributed with a process. The description of the methodology and the implementation of such processes are sometimes documented in dedicated papers, as referred to in the course of this document. The detailed steps of our optimisation pipeline for level 2 data are shown in Figure 3.3. The distributed architecture chosen allows us to execute the pipeline partially, which contributes to the flexibility required for the particularities of each specific level 2 data set and the associated requirements. Similarly, it is also possible to execute the full pipeline following the most appropriate sequence of blocks.
One important aspect to be highlighted regarding the overall approach to this research problem is the joint natural-artificial learning framework based on the interaction of the specialised and multi disciplinary user community with our configurable automatic pipeline. This concept is based on the philosophy of continuous improvement; this means that the results from a particular pipeline block normally feed into the related blocks (see discontinuous lines in Figure 3.3) to refine and improve

their efficiency. In general, the overall result can be retro-fed into the initial catalogue in an iterative process by which the learning process is consolidated.
The design, development and validation methodology have followed a typical V-cycle including Agile techniques for the debugging of the code. In this sense, a set of user requirements and the corresponding verification and validation control process are put in place, and the configuration and overall quality of the system are assessed. In this respect, MATLAB offers an automatic functionality in this area with its *Profiler for Improving Performance* tool.
In terms of the design considerations for the pipeline, the following list of assumptions needs to be taken into account when reviewing the pipeline proposed here. The justification for these assumptions is to be found in the type of images and associated multi-band photometric catalogue that we are using. The distributed architecture of our pipeline can potentially adjust variations on any of these assumptions but the corresponding feasibility study is not part of the scope of this work.

- The astrometric cross-matching is already part of the catalogue. If this were not the case, we would consider the Bayesian astrometric cross-matching described in [Budavári and Szalay, 2008], whose implementation is described in [Marquez, 2012].
- The classification of star/galaxy is already established in a robust manner as part of the catalogue.
- We assumed independence of measurements, and therefore we discarded from this approach those joint measurement frameworks among different surveys or instruments where a correlation might exist between these measurements.
- We assumed that the ratio S/N of the images allows proper differentiating of the background from the foreground in the majority of the cases, especially for bright sources.

3.3 Context of the Problem

This research has merged two disciplines: astronomy and artificial intelligence; AI is the main discipline used in this research and plays the principle role in defining the various methodologies involved, being computer science the frame for the effective implementation of these methodologies, and astronomy is the scientific discipline to which the methodologies are applied. For the sake of external readers who are not necessarily expert in all these disciplines, dedicated bibliographic references for further readings relevant to this research are included as appropriate in this document.

3.4 Knowledge Engineering Framework

As indicated in Section 3.1, we address here the computational level taxonomy, as described in [Mira, 1995], as the main steps followed to create an ontology for the classification of galaxies, considering the lines of research linked to the three objectives presented in previous sections.
This makes use of the discipline of Knowledge Engineering to compile within a framework the semantic and axiomatic relationships established between the different elements of the research.
As done in other sections, we include here a brief description of the context in which we developed our ontology. This comes from the IVOA (International Virtual Observatory Alliance) rules and recommendations and the Web Ontology Language used as a development environment for the ontology. We include here a brief summary of the most relevant aspects used in our ontology development, extracted from [McGuinness et al., 2004] and [Bechhofer, 2009].
We have followed the IVOA Unified Content Descriptor (UCD); therefore the semantics that we used in our ontology is in line with the standard vocabulary for describing astronomical data quantities.

For details about the OWL development tool relevant to its usage for our research here, see [Gray et al., 2008] and [McGuinness et al., 2004] The complexity of the problem presented here requires the use of a level hierarchy as indicated in [Mira, 1995] in order to segment this complexity.
The first level is described through the problem definition using natural language. Therefore, the natural language description of our problem is as follows:

- If a group of photometric measurements for a set of wavelengths can be fit to a Spectral Energy Distribution (SED) function, selected from a exiting library of SEDs, then we conclude that this group of photometric measurements correspond to the same galaxy and its SED follows the one identified from the library.
- The surface brightness of an astronomical source, galaxy in our case, is obtained by summing up the pixel brightness inside a contour. The determination of the contour is done by evolving a curve yielded from a partial differential equation problem for which the termination condition is established as a function of the local gradient of brightness between the center of the source and its immediate surroundings.
- A galaxy Φ is preliminary labelled as isolated for a wavelength λ_i if no other galaxy contour in the neighbourhood is intersecting with the contour of our galaxy under consideration Φ. If this condition is repeated for the galaxy Φ for all its wavelengths then it is labelled as **isolated**. Conversely, a galaxy Φ is contaminating itself if its contour is intersecting any other galaxy contour in the neighbourhood. And a third case would be that the contour of the galaxy Φ under consideration is fully contained in its voronoi cell, thus, it is candidate of isolated galaxy, but at least one other galaxy contour is intersecting with the voronoi cell of our galaxy Φ under consideration. Then, Φ is preliminary labelled as **partially contaminated**.

The second level corresponds to a symbolic level of analysis of the problem, typically named pseudo-code in computer science. This second level for our problem is presented here as follows:

- A Tuple $\overrightarrow{T}$ of n magnitude values in n wavenlengths $\in$ Galaxy Φ?
  - YES: Photometric Bayes Factor is of high value (above a threshold) AND $\overrightarrow{T}$ can fit with at least one SED of a Library of SEDs predefined.
  - NO: Photometric Bayes Factor is below a pre-defined threshold OR $\overrightarrow{T}$ cannot fit with at least one SED of a Library of SEDs
- The active contour AC in iteration step i of a Galaxy Φ in an astronomical image IS defined by the pixels of its surrounding area whose gradient brightness values relative to the center have reached the local maximum value, compared to the next iteration step $i+1$, which correspond to the local boundary of the source. The surface brightness value is obtained from the sum up of the brightness of the pixel within the AC substracting the background brightness for the same area determined by AC.
- The condition of contamination of a Galaxy Φ is labelled with the following rule-based system:
  - A Galaxy Φ_i IS ISOLATED candidate IF its Active Contour $AC_i \in$ its Voronoi Cell AND AC_i IS NOT CONTAMINATED
  - Galaxy Φ_i IS CONTAMINATED candidate IF its Active Contour $AC_i \cap$ its Voronoi Cell $\neq 0$ AND $AC_i \cap$ Voronoi Cell of any other Galaxy $\Phi_j \neq 0$
  - Galaxy Φ_i IS PARTIALLY CONTAMINATED candidate IF its Active Contour $AC_i \in$ its Voronoi Cell AND AC_j of any other Galaxy $\cap$ $AC_i \neq 0$

Finally, the third level corresponds to an implementation level and it involves the selection of the programming language also at the level of software architecture design. This third level is presented in Sections 3.5.2 and 3.5.3

3.4.1 Ontology in Galaxy Classification

The lines of research described in the above sections mainly aim at refining the classification of galaxies in aspects related to photometry. We describe here an Ontology which we have created based on the knowledge acquired in this research. We observe the IVOA rules, recommendations and guidelines for the creation of ontology and associated semantic vocabularies.
From the existing IVOA working groups, described in Chapter 2, the semantics working group provides views on two main areas where this research can be linked: firstly the description of words used in astronomy involving an update of the UCD (Unified Content Descriptors) and the definition of IVOA Standard Vocabulary, and, secondly the creation of an Ontology of Astronomical Object Type - "ObjectTypes099r.owl", see Figure 3.4, which we consider here as the baseline into which we exercised the inclusion of our galaxy classification ontology.
In this research we have expanded the Astronomical Object Type with properties and concepts derived from the main lines of this research.
The concept hierarchy of the IVOA Astronomical Object Type ontology, described in [Cambrésy et al., 2010], is organized around the following top-level concepts which are:

AstrObject

AstroPortion

AtomicElements

EMSpectrumRange

Measurement

Morphology

Process

These sections can be split in two categories: AstrObject and AstroPortion subsume the astronomical object types and their constituents, while the other sections are ranges of properties used to define the concepts of the AstrObject and AstroPortion sections.
For our research we will adopt the same colour-coding used in the OWL ontology of Astronomical Object Types which is defined in [Cambrésy et al., 2010] as follows:

- Yellow: concepts for which we have necessary conditions but no definition
- Orange: concepts for which we have at least one definition
- White: concepts from other branches than the one considered, which can be either defined or not, but have been coloured white to enhance legibility.

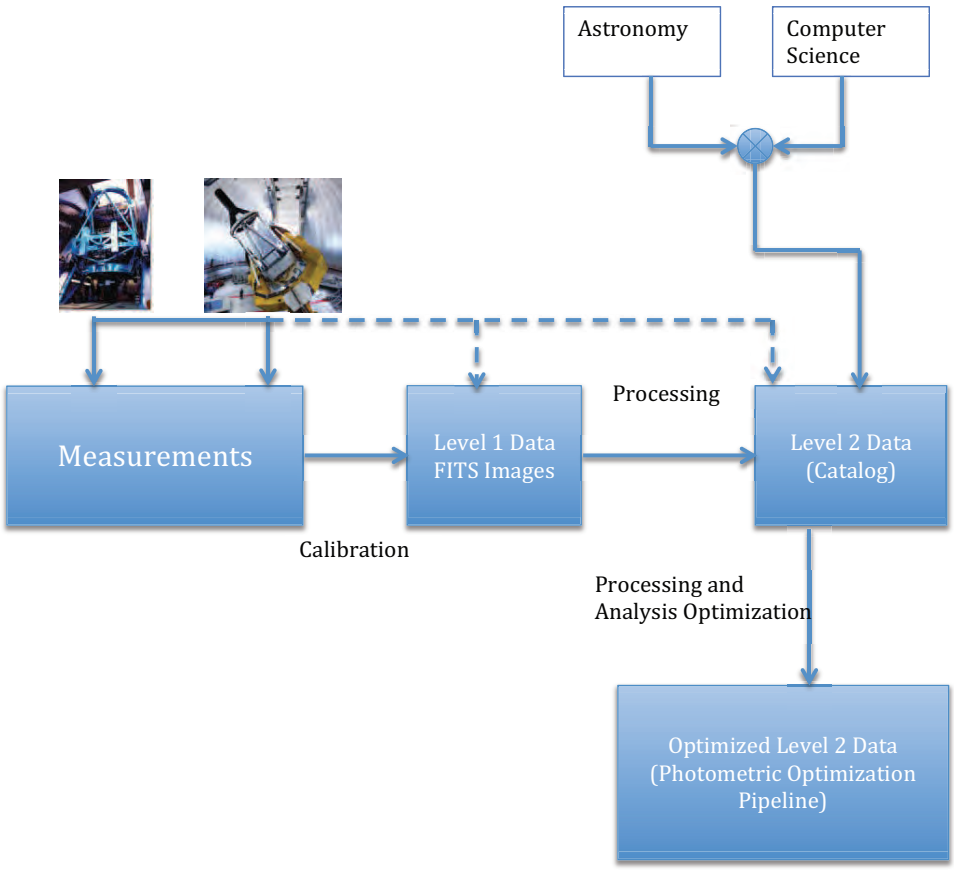

Figure 3.1: This figure shows the block of the main generic steps and level of information involved in astronomical data processing and analysis.

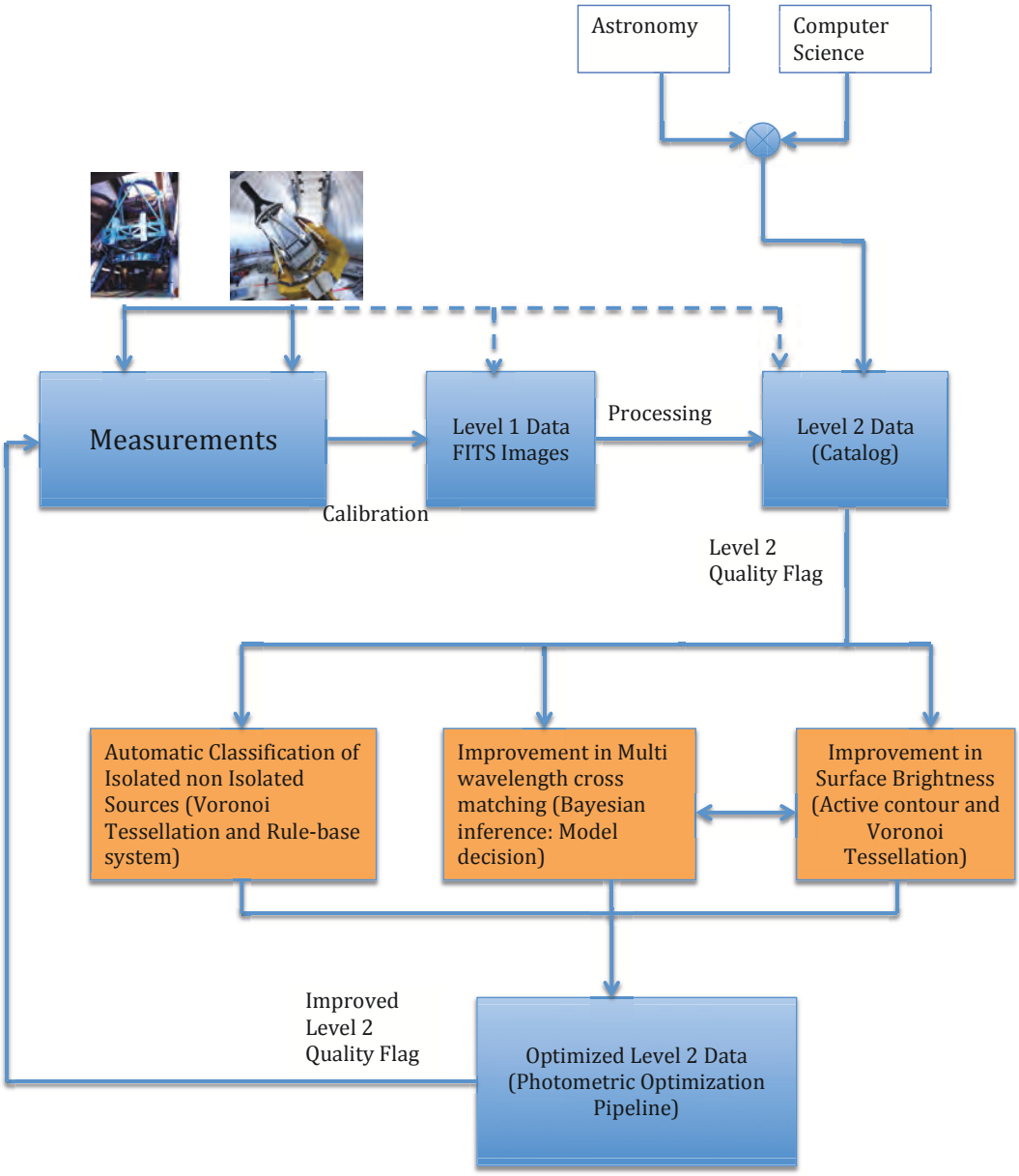

Figure 3.2: Compared to Figure 3.1 this figure shows the ingestion points where our optimisation pipeline can run. These optimisation modules are indicated in orange colour.

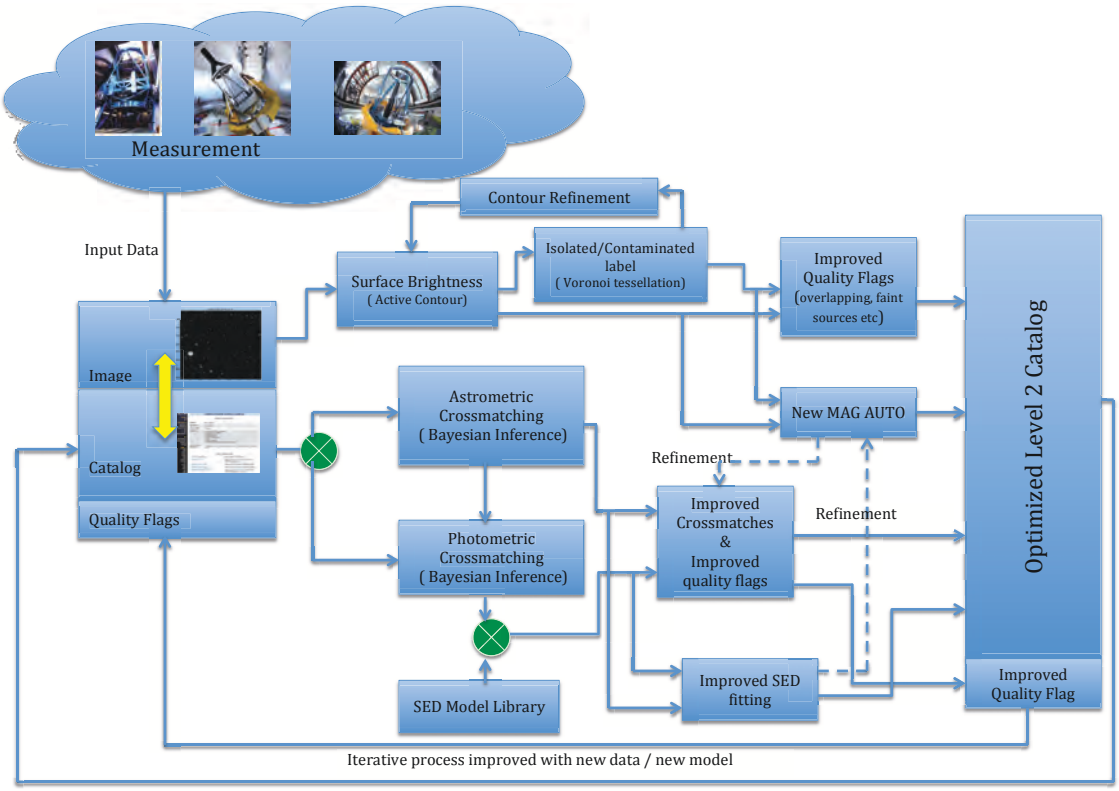

Figure 3.3: This figure shows the details of the main components of the optimisation pipeline for the level 2 data.

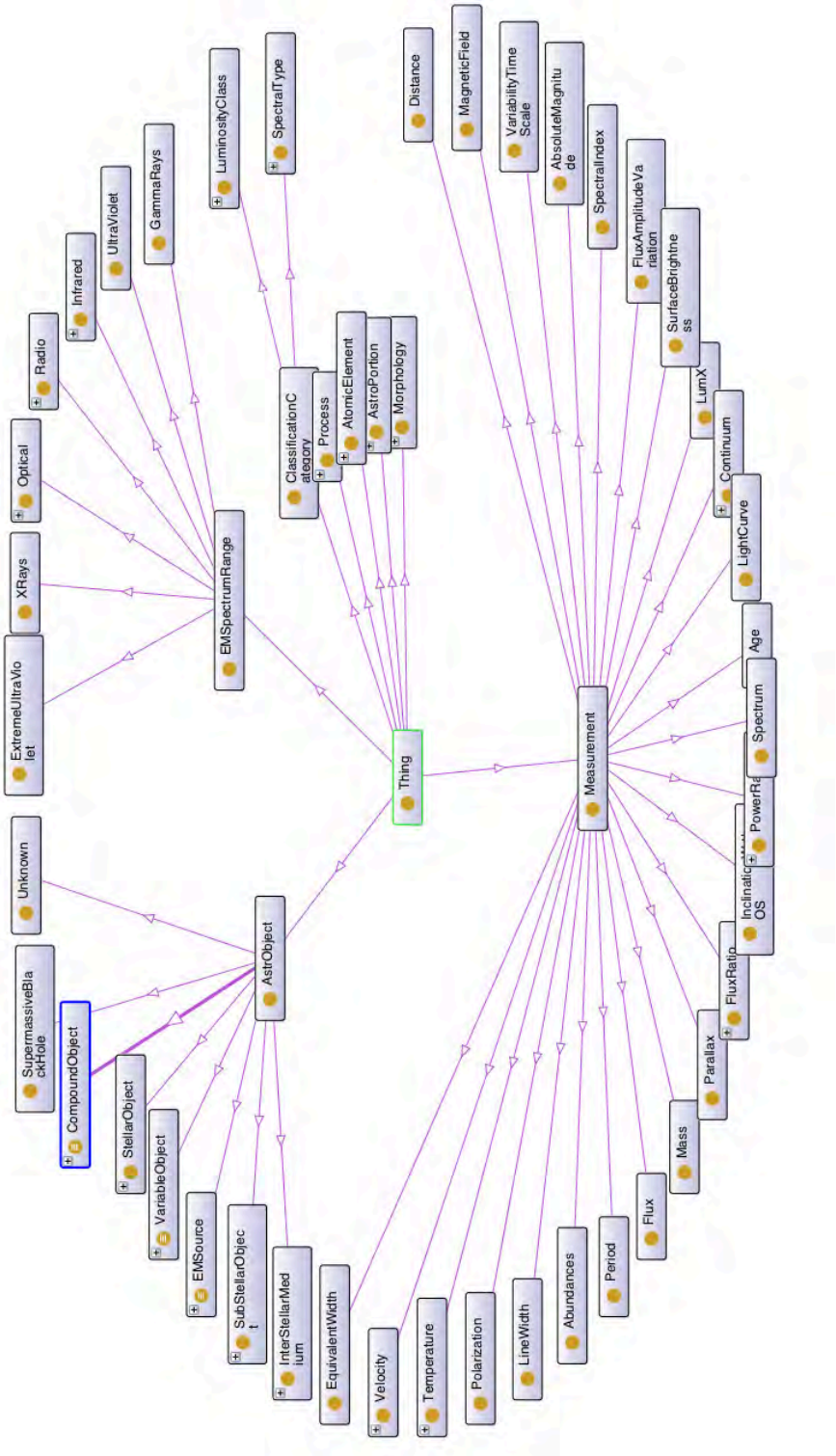

Figure 3.4: Ontology of the IVOA Astronomical Object Type defined by the IVOA Semantics WG

Initially we created an Ontology for this research observing the general rules and recommendations published by IVO in this area. Figures 3.5 and 3.6 are included in an on-line platform for readability purposes. Figure 3.6 shows the result of merging the Astronomical Object Type Ontology of IVOA with our Ontology for this research. In this merging exercise we did not implement any adaptation to the naming of classes or properties or to hierarchies and usage from the original ontology.
In terms of definition of properties, [Cambrésy et al., 2010] already introduced the *hasConstituent*, *hasComponent* and *hasPortion* properties (see [Cambrésy et al., 2010]). Other properties were introduced to describe astronomical objects via not only their constituents but also their emission, the processes they are subject to, the measurements made and the morphological features of their spectral characteristics.
Table 3.1 describes the list of most relevant properties included in the IVOA astronomical object types. And Table 3.2 indicates the most relevant aspects of the properties used in the Ontology of this research being merged with the IVOA Astronomical Object Types one.

3.5 Development Methodology of the GCES

The fulfilment of the objectives indicated in the previous section along with the context in which the problem is stated required the consideration of a complete software development life cycle. In terms of software development philosophy, a hybrid between a V-cycle and an Agile methodology has been considered in order to extract the maximum benefit from both approaches.
The justification for this hybrid approach can be found in the complexity and context of the problem, which required on one hand the initial establishment of a solid core set of user requirements and calibrated input data, and on the other hand the flexibility to perform quick loops of upgrades and corrections in the part of the code based on performance and readiness assessment from partial results.

3.5.1 User Requirements

In order to establish the appropriate scope of the problem described here and with it the resolution proposed, a set of requirements have been identified which detail to an adequate level of granularity the sub-problems to be resolved.
Pipeline.SYS010. The system shall identify isolated sources following unambiguous rules.
Pipeline.SYS020. The system shall implement the cross-matching of multi-wavelength sources by using photometry
Pipeline.SYS030. The system shall provide surface brightness values of extended astronomical sources based on the pixel brightness computation within an optimised contour of such sources.
Pipeline.PERF010. The system shall allow expanding capability with the provision of relevant information.
Pipeline.PERF020. The software for the generation of random values shall provide enough confidence of these random capabilities.
Pipeline.PERF030. All standalone executions of the pipeline shall provide representative results with a running time of less than 10 days.
Pipeline.USR010. The user shall be capable of partially or totally running the pipeline in a configurable manner.
Pipeline.USR020. The user shall be able to input the predefined by design configuration information in a simple manner.
Pipeline.USR030. The usage of each software module shall be documented such that any user can understand and use it.
Pipeline.VER010. The software developed shall be verified against these requirements.

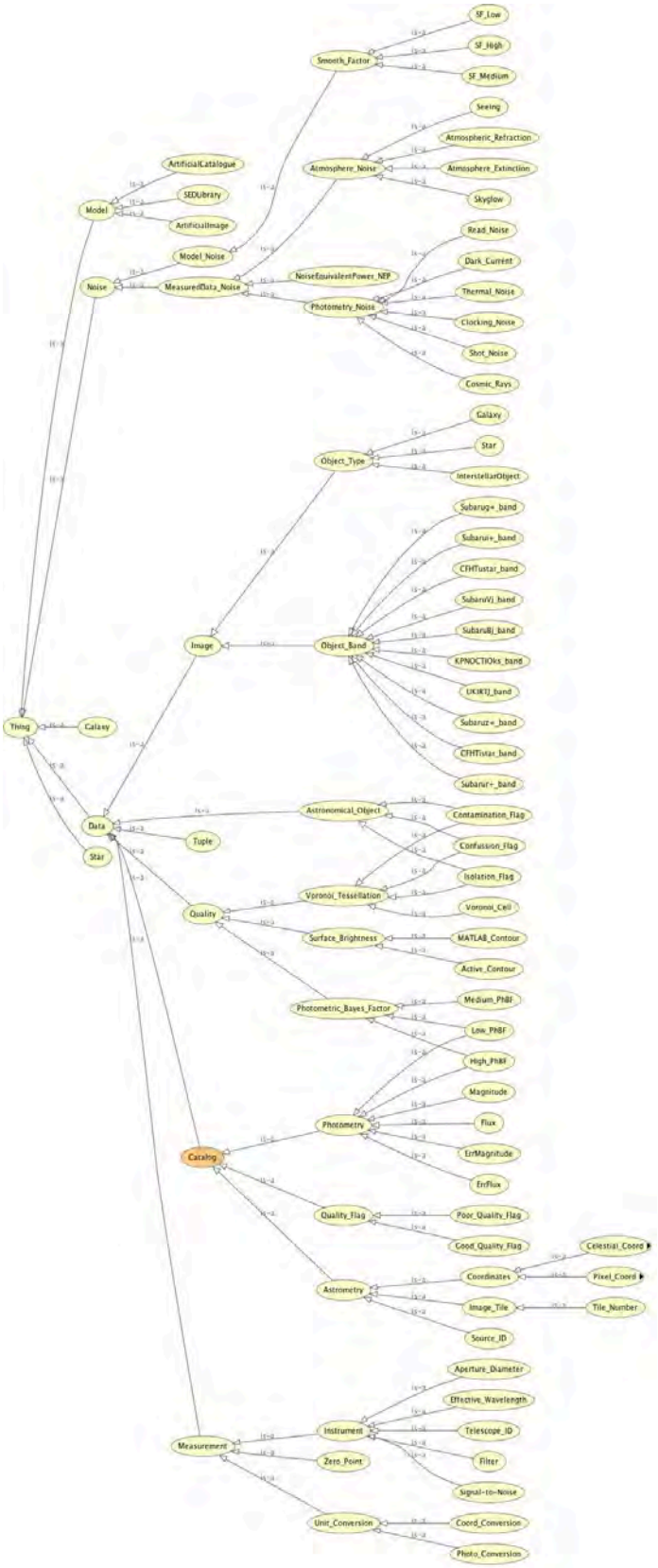

Figure 3.5: Inferred ontology of our research

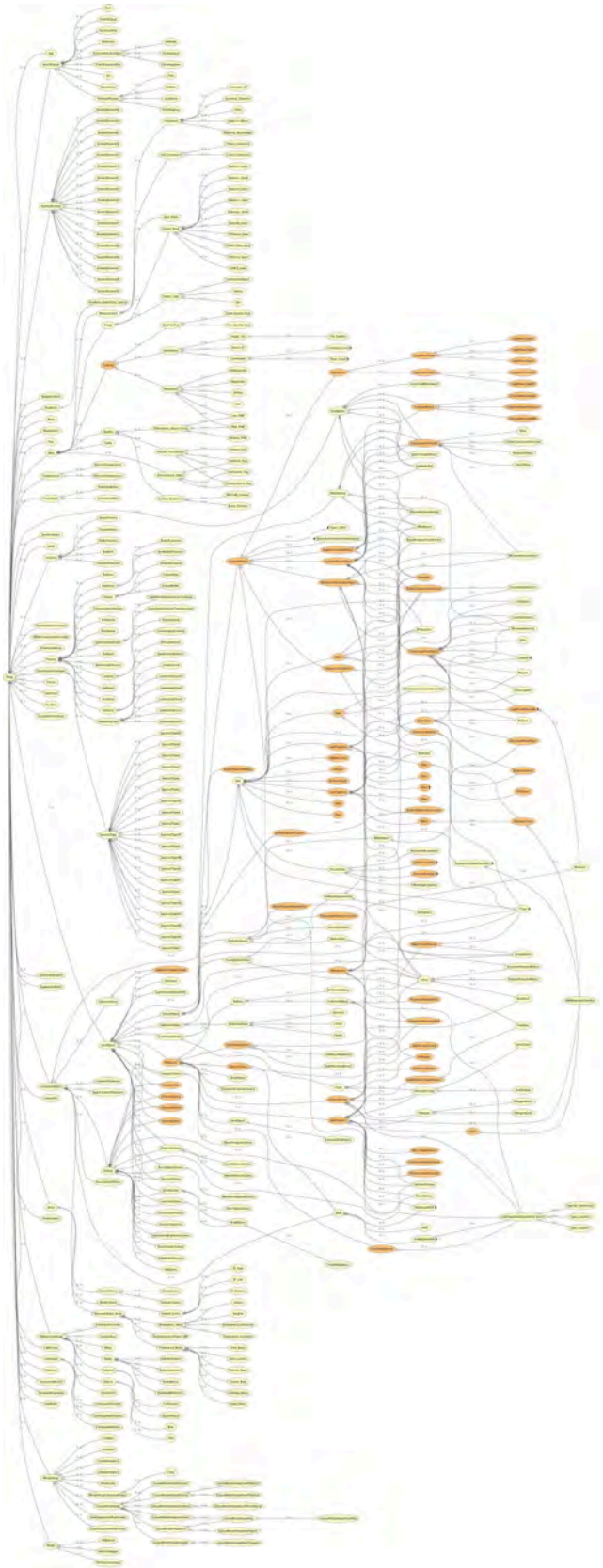

Figure 3.6: Inferred merged ontology of our research with IVOA Ontology on astronomical object types

Table 3.1: Properties of IVOA astronomical object types ontology.

Name	Domain	Range
hasEmissionIn	AstrObject	EMSpectrumRange
hasPeakEmissionIn	AstrObject	EMSpectrumRange
hasMeasurement	AstrObject	Measurement
hasHighMeasurement	AstrObject	Measurement
hasLowMeasurement	AstrObject	Measurement
isMeasuredFor	Measurement	AstrObject
hasProcess	AstrObject	Process
hasVariabilityProcess	VariableObject	ProcessVariability
hasBurstProcess	VariableObject	Explosion
hasPeriodicProcess	VariableObject	Eclipse OR Rotation OR Pulsation
hasTransientProcess	VariableObject	ProcessVariability
hasPortion	AstrObject OR AstroPortion	AstroPortion
isMorphologyOf	Morphology	AstrObject
hasMorphologyOf	AstrObject	Morphology
hasConstituent	CompoundObject OR AstroPortion	AstrObject
hasComponent	CompoundObject	AstrObject
isConstituentOf	AstrObject	CompoundObject OR AstroPortion
isComponent	AstrObject	CompoundObject
hasSpectralLine	AstrObject	AtomicElement
hasEmissionSpectralLine	AstrObject	AtomicElement
hasAbsorptionSpectralLine	AstrObject	AtomicElement
hasVariabilityTimeScale	VariableObject	VariabilityTimeScale
isPortionOf	AstroPortion	AstrObject OR AstroPortion

Table 3.2: Properties of the ontology developed in this research

Name	Domain	Range
hasBadFit	Catalog	Photometry
hasBadMatch	Photometric_Bayes_Factor	Noise
hasBayesFactor	Quality	High_PhBF or Low_PhBF
hasComponentOf	Data or Model	Measurement
hasGoodFit	Good_Quality_Flag	Measurement
hasGoodMatch	Good_Quality_Flag	High_PhBF
hasHighQuality	Good_Quality_Flag	High_PhBF or Voronoi_Tessellation and Surface Brightness
hasIsolatedFlag	Isolation_Flag	Quality or Model
hasLowQuality	Poor_Quality_Flag	Measurement or Noise
hasNonIsolatedFlag	Contamination_Flag	Quality or Model
hasObjectSpectre	Data	Measurement and Image
hasTupleOf	Catalog	Measurement
hasVoronoiCell	Model	Image
InsideOf	Confusion_Flag or Contamination_Flag or Isolation_Flag	Model or Image
isComposedOf	Data or Model	
OutsideOf	Confusion_Flag or Contamination_Flag or Isolation_Flag	Model or Image
OverlappingWidth	Contamination_Flag or Confusion_Flag	Voronoi_Cell

Pipeline.VAL010. The software developed shall be functionally validated against the objectives of this research under nominal and degraded scenarios.
Pipeline.CM010. The software shall be implemented, validated and documented under general configuration management rules allowing the unambiguity, trace-ability and reproduce-ability of the software.
Pipeline.QA010. The software shall be inspected and audited under general quality rules applicable to the development platform selected.

3.5.2 Software Architecture Design

As indicated in [Rumbaugh et al., 2004], a software-intensive system demands modelling from a number of perspectives as a central part of all the activities that lead up to the development of good software. With models we communicate the structure and behaviour of our system and we visualize and control the system's architecture. This information allows us to identify opportunities for optimisation and to manage the associated risks.
The experience of modelling in engineering disciplines suggests four basic principles:

- The choice of what models to create has a profound influence on how we tackle a problem and how a solution is shaped;
- The level of precision in the definition of each model may vary;
- Reality and best models are connected;
- The consideration of multiple viewpoints through simple independent models is the best approach towards modelling.

For the modelling of our pipeline we have chosen the use of UML (Unified Modelling Language) which allows us to specify, construct, visualize and document the software system of our processing pipeline.
Among the different diagrams in UML and considering the coding approach used, we choose the use case, packaging, class/object and activity diagrams as the most appropriate ones to represent the model of our software architecture.
The use case view diagram of a system encompasses the use cases which describe the system behaviour from the users perspective.
We construct a use case of our pipeline system to represent a set of sequences which describes the intended behaviour of our system according to the requirements established in Section 3.5.1. This is one of the diagrams which models the dynamic aspects of our system.
The diagram in Figure 3.7 represents the main use cases of our pipeline from a user point of view. In this diagram we will show only those uses cases which are important for understanding the behaviour of the system or the part of the system in its context. We will also show only those actors which relate to these use cases.
In this use case diagram we have identified different kind of users, which forms a coordinated and well-structured user community. Each type of user retains a specific interest and context where the usage is deployed. The science user follows the pursuit of "doing Science", and the pipeline developed here is considered a possible tool to allow this user the fulfilment of this ultimate goal. In order to do so, the support from what it is called the "Astrostatistics Work group" is very important, because this work group consolidates the knowledge of the two disciplines involved in this research: astronomy and computer science. In this sense, level 1 and level 2 experts are part of this work group incorporating sound expertise in the area of instrument-related aspects and scientifically meaningful algorithm aspects. Finally, the computer science expert contributes with expertise by computing all the data and related algorithms, making the best use of available technologies and anticipating the

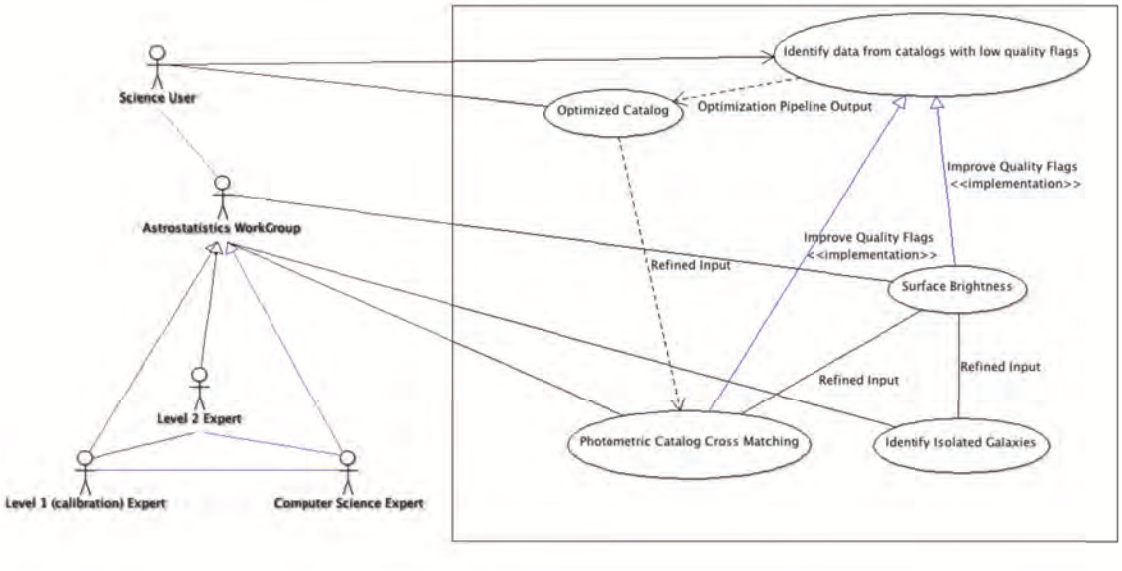

Figure 3.7: Use Case

suitability of infrastructures for future technologies.
Only properties of the classes which are important for understanding the abstraction in its context will be represented in the diagrams.
The design view diagram, which includes the object and activity diagrams, supports the functional requirements of the system, that is the services that the system should provide for its end users. The object diagram captures the static aspects of the system, and the activity diagram captures the dynamic aspects.
In our UML diagrams we used the notion of system indicated in [Rumbaugh et al., 2004] as a collection of subsystems organized to accomplish a purpose and described by a set of models, possibly from different view points.
We use the class diagrams here to model the static design view of our pipeline system.
In the class diagrams in Figures 3.8 to 3.10, the dashed directed lines represent dependencies stating that one object (coded with script of function) is using the information and services of another object. When one extreme of the line contains an unfilled diamond, this means that we are using an aggregation relationship. Similarly, generalizations represents different levels of abstraction and these relationships are represented by a continuous line ending in an unfilled triangle.
In the class diagram for the photometric multi-wavelength cross-matching, we observe that there are four main aspects considered in the software development: the data, the model, the processing of the data and the configuration capabilities. Regarding the data, a formatting step was included in the pipeline in order to work with comparable well-calibrated data.

3.5.3 Software Development

The main reason for using MATLAB for this implementation was the need to efficiently focus on the mathematical problem associated with each objective. In that respect, MATLAB presented a solid effective and quick basis for developing the code, then integrating and validating it. In addition to that, the versatility of MATLAB in terms of data types and quick prototyping was

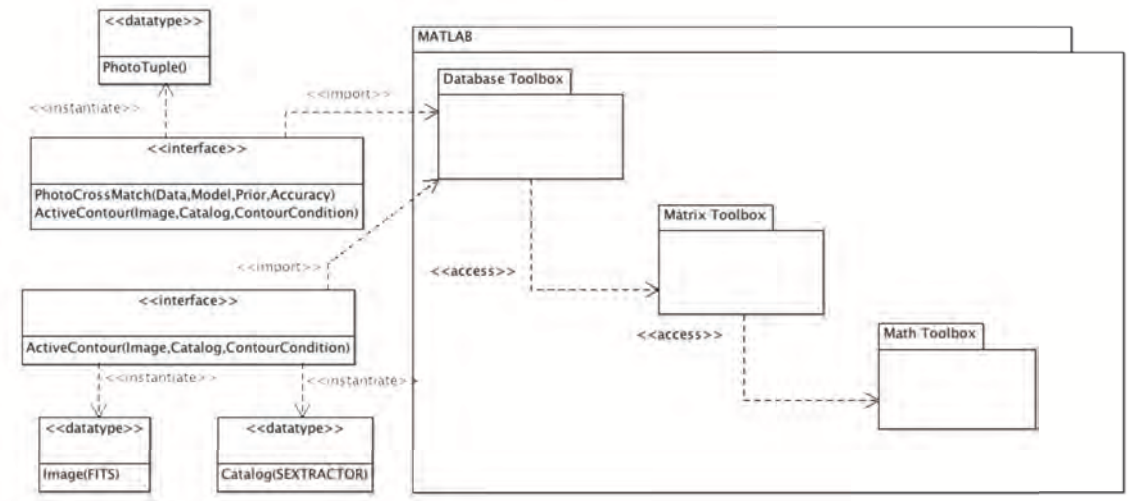

Figure 3.8: Packaging model for the pipeline

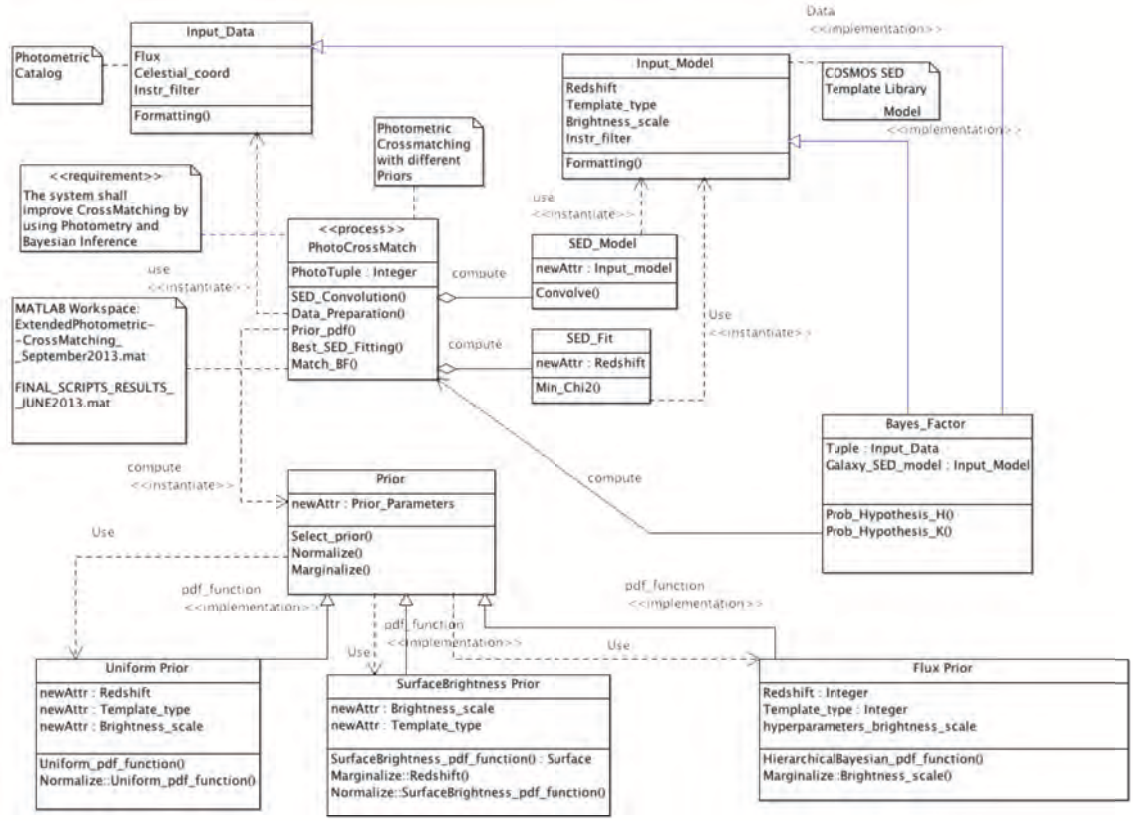

Figure 3.9: Class model for the photometric cross-matching

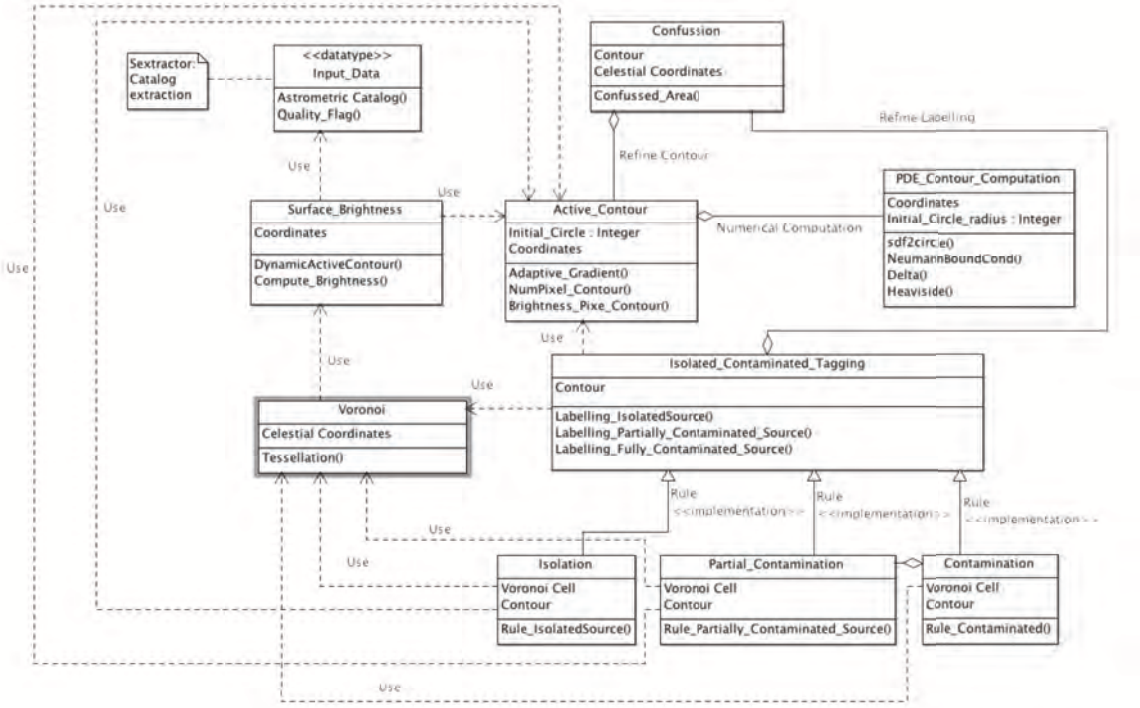

Figure 3.10: Class model for the surface brightness computation by using active contour. This model also represents the isolated versus non-isolated labelling by using Voronoi tessellation

considered a strong point in selecting that platform. Finally, the parallel computing capabilities of MATLAB in terms of multiple parallel workers, GPU usage and vectorisation allowed an initial quick assessment of this feature by easily exporting a small part of the code to a parallel computing framework and establishing a benchmark with the non-parallel computing framework, yielding promising results. However, this work will not be considered part of the scope of this research.
In summary, this pipeline is composed of three main branches, each of them using specific methodologies. The different steps of this pipeline contribute to the improvement of knowledge engineering in the context of this problem. This means that the improved catalogue of galaxies can be linked to knowledge engineering capabilities potentially expandable to other surveys; for example, the criteria used in the identification of isolated versus overlapped galaxy tagging can be formally exported to a platform such as IVOA (International Virtual Observatory Alliance) where an exhaustive astronomical taxonomy has been deployed. Similarly, we can define interfaces with these kinds of platforms for each of the criteria used in this pipeline.
As explained in the above sections, the pipeline that we have designed is composed of three main blocks, each of them supported by specific artificial intelligence methods whose outputs can be considered as input to the relevant knowledge engineering frameworks. These three blocks are:

- improvement of cross-matching and best SED fitting by using a new photometric probabilistic Bayesian Inference which belongs to the Data Mining domain. From this line of the research the class, relevant properties and inferred axioms related to photometric cross-matching multiband sources were included in our ontology.
- galaxy surface brightness computation by using a new active contours algorithm which belongs to the the artificial vision domain. The class, relevant properties associated to the galaxy surface brightness and inferred axioms were also included in our ontology.
- isolated versus contaminated galaxy tagging by using a simple rule-based system, based on the contours obtained in the previous block, which belongs to the knowledge engineering domain. The class, associated properties related to the classification of isolated, contaminated and blended astronomical sources and the inferred axioms were part of our ontology.

An important aspect of the overall pipeline design are the interfaces between the different software modules that allow us to establish the proper data flow across the full pipeline. The structured approach is based on MATLAB independent functions whose input and output parameters are the connection points of these interfaces.
Each main branch of this pipeline has been independently validated. The results obtained from real data have been thoroughly analysed to determine the added value offered by the contributions presented in this research. Details of this assessment are provided in Chapter 5.

3.5.4 Verification and Validation

This section addresses the approach followed in the integration, verification and validation of the different software modules that compose the pipeline design of this research.
As indicated in [Rumbaugh et al., 2004], use case diagrams are important for testing executable systems through forward engineering. The notion of usage of the pipeline through the definition of use cases allows us to establish the necessary functional chains which validate the full pipeline in terms of the overall data processing functionality. Using this it is also possible to assess the performance of the design in terms of the accuracy of the results and computational costs.
The following table includes the typical structure and information of a verification control document, where the first column is the requirement identification, the second column is the result - PASS, FAIL- , the third column indicates the verification method - analysis or test, and finally the last

column details the evidence through which the verification result is identified.

Requirement ID	Result	Method	Evidence
Pipeline.SYS010	PASS	Code analysis and test	See Chapter 4
Pipeline.SYS020	PASS	Test results	See Chapter 4
Pipeline.SYS030	PASS	Test result	See Chapter 4
Pipeline.PERF010	PASS	Code Analysis	See code
Pipeline.PERF020	PASS	Test	See Chapter 4
Pipeline.PERF030	PASS	Test	See results in Chapter 4
Pipeline.USR010	PASS	Analysis	See code documentation
Pipeline.USR020	PASS	Analysis	See code documentation
Pipeline.USR030	PASS	Analysis	See code documentation
Pipeline.VER010	PASS	Test and Analysis	See code and results in chapter 3
Pipeline.VAL010	PASS	Test	See results in Chapter 3 and 4
Pipeline.CM010	PASS	Analysis	See 3.5.5
Pipeline.QA010	PASS	Analysis	See 3.5.5

Table 3.3: Pipeline Verification Matrix

3.5.5 Quality Assurance and Configuration Management

Regarding code quality, MATLAB includes a functionality named profiler which allows it to perform automatic review of the code. The result is a report which shows the level of adherence of the code under analysis to good practices. Three main aspects are considered in this analysis:

- General coding-writing good-practices: these are a set of rules common to the majority of programming languages. E.g., to include comments in the header of each function or software module; to avoid global variables, etc.
- Performance-time: these rules are more intrinsic to the internal features of MATLAB in terms of computational time.
- Performance-memory: these rules points to increase performance related to the move of data across memory blocks. This is also connected with the computational speed and these considerations are strongly linked to the internal features of MATLAB.

More detailed information is provided in [Altman, 2014] and [McKeeman, 2008].
Thus, the code produced in the context of this research has gone through the MATLAB profiler audit and the outcomes of the report were considered to obtain the final improved versions of the software. These reports will be attached in electronic format to this document.

Regarding Configuration Management the configuration status of each piece of code produced has been kept under version control and documented in a live word document, named Software Configuration Items Delivery List (SCIDL). The changes implemented in the code are declared in the SCIDL and also in the header of the MATLAB scripts.

3.6 Cross-Matching of Multi-band Galaxies

The statistical inference provides a means for assessing the plausibility of one or more competing models (also named hypotheses), and estimating the model parameters and their uncertainties, which is part of the data analysis discipline.
The model selection problem in Bayesian inference resolves into which of two or more competing models is most probable based on our state of knowledge at the moment of the analysis. In model selection, we are interested in the most probable model, independent of the model parameters. The hypothesis space of interest is discrete (although its parameters may be continuous). The models may differ in form or number of parameters. In this case, the Bayes' theorem is used to compare competing models by calculating the probability of each model as a whole. Often it is useful to consider the ratio of probabilities of two models, rather than the probabilities directly. Assuming that we have no information leading to a prior preference for one model over another, the prior odds ratio will be unity, and the odds ratio will equal the Bayes Factor, i.e., the ratio of Bayesian evidence of each model.
Evidence refers in this document to the probability of the data conditioned on the hypothesis, i.e. $p(D|H)$ and $p(D|K)$ will refer to the evidences of hypotheses H and K respectively.
As indicated in [Trotta, 2008], a crucial consequence of the marginalization procedure used to calculate the Bayesian evidence is that the Bayes Factor automatically favours simpler models unless the data justify the complexity of more complicated alternatives; this means that the Bayes Factor balances quality of fit versus extra model complexity, rewarding highly predicting models, and penalizing "wasted" parameter space.
The probability evidence for a given model is therefore determined principally by the mean value of the resulting likelihood function (averaged over the prior probability density function (pdf)).
Applying Bayes' theorem to the ratio of the posterior pdf of two models, A and B, we obtain the following:

$$\frac{p(A|D,I)}{p(B|D,I)} = \frac{p(D|A,I)}{p(D|B,I)} \cdot \frac{p(A|I)}{p(B|I)} \tag{3.1}$$

Where A and B represent two propositions asserting the truth of the two models or hypotheses of interest; D is the proposition representing the data and I is the proposition representing our prior information.
If there is no specific preliminary information in favour of one model or the other, we might take the ratio $\frac{p(A|I)}{p(B|I)} = 1$

3.6.1 Photometric Bayesian Inference

As indicated in [Marquez et al., 2014], the problem of identifying the same astronomical source in different exposures and across instruments has long been studied for its statistical and computational complexity. This becomes challenging when faint sources are part of the data and/or very large catalogues are considered; manual or automatic operations based on general thresholds are not only too tedious but also too inefficient to be considered an option for large survey astronomical data. Typically, when trying to determine whether multiple sources do actually belong to the same object, the initial information considered, are the coordinates of the sources in the sky, but highly-populated areas often yield degenerate cases when multiple possible matching configurations have similar likelihoods. Bayesian inference can alleviate this problem by including photometric measurements. The Spectral Energy Distributions (SEDs) of candidate associations are compared to the models to check the photometric evidence for a good match. In the following paragraphs we will describe the implementation, and in Chapters 4 and 5 a discussion of the preliminary results from simulated data and from the COSMOS catalogue data are presented. The main conclusions

of this approach are included in Chapter 6. The context of the problem described here is the photometric cross-matching problem, primarily for galaxies. The target is to design and implement an automatic pipeline capable of reliably identifying cross-matches. The methodology applied is the Bayesian inference for the hypotheses decision problem, where the SED built from measured data is compared to SED template library models to see which hypothesis (matching or non-matching) is most plausible. The plausibility of these two hypotheses will be assessed through the computation of the ratio of Bayesian evidences of each of the hypotheses, which gives the Bayes Factor. This method is well justified for this problem where large training sets are not usually available. In [Budavári and Szalay, 2008] an astrometric cross-matching problem was addressed by comparing the likelihood of two complementary hypotheses: hypothesis H represents the case in which the astrometric positions of a group of detections of sources of different bands correspond to the same object, and hypothesis K stands for the case in which detections of multi-wavelength sources do not belong to the same object. For Gaussian (or Fisher) uncertainties, the astrometric Bayes Factor can be analytically calculated and efficiently implemented for the N-way case using recursions.
We are now focusing on the photometric cross-matching aspect of the problem, and therefore fluxes of sources become relevant information. Here we apply a similar methodology, only in this case instead of identifying matched sources based solely on their positions, their flux measurements in different wavelengths will yield the plausibility of being or not the same source. The fact is also highlighted that uncertainties related to photometric measurements are a challenge today in the astronomical field, being strongly linked to the physical capabilities of the associated hardware. [Marquez and Sarro, 2013] describes how data can be refined through improvement of the setup of the instruments and optimisation of uncertainties (though the uncertainties cannot be reduced to zero). The approach proposed in [Marquez et al., 2014] also implements galaxy SED fitting and, with this, an assessment of the completeness and suitability of the SED template library is possible. In this research we consider as candidates the resulting associations from astrometric matches and compare H to K in light of photometric measurements.
Two hypotheses, H and K, define respectively the cases of data, being a multi-wavelength tuple of source fluxes along with their uncertainties and errors - as for independent observations and measurements, the uncertainties and errors relevant for each measurement process will also be independent, corresponding to the same source, detected in different bands, or the case of data corresponding to different sources, each one in a different band.
We assume independence of measurements, and therefore we discard from this approach joint measurements among different surveys or instruments where a correlation might exist between these measurements.
Therefore, in this case, two complementary hypotheses, H and K, for the photometric data can be described as:

- Hypothesis H: all the measured flux of the tuple j under analysis correspond to the same source.
- Hypothesis K: not all the measured flux of the tuple j under analysis correspond to the same source.

In Bayesian inference, the term evidence refers to the product of the likelihood and the prior integrated over all the model parameters. The probability of the data given the model and parameters is the likelihood, and it can be obtained for all models as the pre-computing step in a SED fitting problem which is detailed here later. The ratio of evidences is named the Bayes Factor. In this case, we compute the photometric Bayes Factor B_{ph} as the ratio between the two evidences $p(D|H)$ and $p(D|K)$, based on the hypotheses H and K. This will help to identify which of the two hypotheses described above is the most plausible one in the probabilistic scenario.

Table 3.4: Filters

Filters			
CFHT u* band	Subaru Bj band	Subaru Vj band	Subaru g+ band
Subaru r+ band	Subaru i+ band	Subaru z+ band	CFHT Ks band
CFHT i* band	UKIRT J band		

$$B_{ph} = \frac{p(D|H)}{p(D|K)}, \tag{3.2}$$

where:

- **Data (D)**: tuples of fluxes in the n bandwidths under study, $D = \vec{f}_i, i = 1, n$
- **Model(M:H,K)**: set of SED templates, convolved with the respective band filters.

When the Bayes Factor is significantly larger than unity, the data have supporting evidence for a true association, but low values argue for no match. When the Bayes Factor is 1, the data do not provide information with which we can decide for one of the hypotheses. To evaluate this ratio we marginalize over a familiar likelihood function. This threshold is known by the Jeffrey's scale defined in [Jeffreys, 1998].

3.6.2 Description of the Models

In order to create a realistic model library of spectral energy distributions across a range of redshifts for the problem of the photometric cross-matching, we have used the Le PHARE package created by Stèphane Arnouts and Olivier Ilbert in the CFHT and Laboratoire d'Astrophysique de Marseille and the "Photoz" cpp library created by Dr. Budavari [Budavári et al., 2000]. Le PHARE is a set of Fortran commands to compute photometric redshifts and to perform SED fitting. Photoz is a library of cpp modules which allows the user to perform, amongst other functions, many useful operations with redshifts and SED models. More details of the Le PHARE package are provided in [Arnouts and Ilbert, 2011].
A library of SEDs models is built upon the normalised convolution of the galaxy spectrum templates with the response function of the measurement systems, which includes the filter transmission, quantum efficiency, optical effect of the instrument, etc, corresponding to the bands under consideration. The spectrum of the galaxy is defined in terms of a restframe spectrum of a galaxy at different redshifts. This yields a transmitted flux at an effective wavelength, as indicated in [Walcher et al., 2011] (for more details read [Fukugita et al., 1995]). The use of appropriate SED templates becomes relevant in terms of precision and coverage of the model against the data. These types of models are used for photometric redshift estimation and for SED fitting, as described in [Walcher et al., 2011], where the construction of galaxy SEDs is identified as a complex problem.
Therefore, we firstly select the SED libraries and filter bands adequate for COSMOS by selecting them from Le Phare repository and we compute the theoretical magnitudes, by using Le Phare commands. Then, we convolve the obtained SED magnitudes with a set of selected filters shifted along a specified range of redshift (z), using the photoz++ package of Budavari as presented in [Budavári et al., 2000]. This allows us to build a multidimensional grid of SEDs.
The practical steps followed for the creation of the grid, referred to later as G, is described in Chapter 4.

For our research the SED library built includes 31 SEDs used in [Arnouts and Ilbert, 2011] and [Ilbert et al., 2009] for COSMOS photo-z.

Following the proposed parameters for the SED modelling in [Budavári and Szalay, 2008], the SED model depends on the the template number T, the redshift z and an overall brightness scaling factor α, which can be defined as a function of the apparent magnitude as we will see below. Therefore the SED model can be expressed as $M = \{\alpha\vec{f}(T, z)\}$ where T is the template type and gives information about the shape of the spectra when it was emitted; z is the redshift and α is the overall brightness scale factor which allows comparison of brightness measurements from SED models in different bands. This factor has been implemented by defining a band as reference and then computing the brightness scale factor of the other bands against the reference one. For example, if we select band R as the reference band, then the scale factor is computed as:

$$\alpha = \frac{f_R}{G_R(T, z)} \tag{3.3}$$

where f_R is the flux for a magnitude in band R, and $G_R(T, z)$ is the transmitted flux of an SED model in band R for a template type T and a redshift z. The value of α is defined for a range of f_R values which correspond to a range of magnitudes m covering the maximum and minimum values of data magnitudes.
Mathematically, this model will be represented, for each band, as a grid of three parameters and each cell of this grid will propose a value of the model to be compared with the data in order to compute the likelihood, as we will see below.
The SED grid can be represented as a matrix of the same number of dimensions as parameters considered, this means, $G(T, z, m)$. A band, R in this case, is selected as the reference band and then the scale factor is expressed in terms of:

$$\alpha = \frac{10^{0.4(m-m_R)}}{G(T, z, m)} \tag{3.4}$$

Where m_R is the apparent magnitude in the R band.
Fitting the SED of galaxies uses semi-analytic models, for instance that of [Bruzual and Charlot, 2003], or empirical template spectra such as those reported in [Coleman et al., 1980]. In the simplest case, a discrete number of templates are used which are scaled and redshifted to match the measured broadband fluxes.
For a given T template, z redshift, and α brightness, we can evaluate the simulated fluxes in any photometric system, which can then be compared with the $\{g_i\}$ observations. Typically

$$\chi^2(\alpha, T, z) = \sum_i^n \frac{1}{\sigma_i^2}\Big[g_i - \alpha\, f_i(T, z)\Big]^2 \tag{3.5}$$

is minimized to find the best-fitting model, which is equivalent to the Maximum-Likelihood Estimation (MLE) with a Gaussian,

$$\mathcal{L}(\alpha, T, z) \equiv p\left(\{g_i\} \,|\, \alpha, T, z\right) = \frac{1}{Z}\exp\left[-\frac{\chi^2(\alpha, T, z)}{2}\right], \tag{3.6}$$

where Z is the normalization constant which depends on the covariance matrix, or in this case, the $\{\sigma_i\}$ diagonal elements. As shown in Chapter 5, the models used fit real COSMOS catalogue data considerably well, whereas for artificially non-matched data the fit to the models is clearly poor .
After computing the probability of the data given the model parameters (for all parameters), the one with the lowest χ^2 yields the best SED fit.
The Bayesian fitting method described in [Walcher et al., 2011] involves a pre-computing of the probability of the data given the model parameters for a discrete set of models SEDs. In our case, a similar pre-computing step is carried out in order to obtain the ratio of Bayesian evidences called

the Bayes factor.
The computation of this likelihood allows for an immediate identification of the best fit of the data, tuple of fluxes, against a specific SED model, being a tuple of fluxes of a specific redshift and type of galaxy selected from the grid of parameters. The larger the mismatch, the lower the corresponding likelihood. The photometric accuracy can be modelled in terms of a Gaussian distribution as indicated in [Budavári and Szalay, 2008]. Therefore we can write the likelihood distribution as follows:

$$p_i(\vec{g_i}|\vec{\nu}(z,\alpha,T),H) \propto e^{\frac{-\chi^2}{2}} \qquad (3.7)$$
$$\chi^2 \propto \sum_i^n \frac{\left(\vec{g_i} - \alpha\vec{f_i}(T,z)\right)^2}{\sigma_i^2}$$

where $p_i(\vec{g_i}|\vec{\nu}(z,\alpha,T),H)$ is the probability density function of measuring $\vec{g_i}$ fluxes for an object i with properties given by a spectral energy distribution model $\alpha\vec{f_i}(T,z)$. This likelihood function is also used in the photo-z methods. Once the probability of the data given the model has been computed for all models, the one with the lowest χ^2 yields the best SED fit. Figure 5.3 shows different types of results of SED fitting for blue and red galaxies from COSMOS data sets (see details about this data set in Chapter 5). The two figures at the top, Fig a and Fig b, represent a good case of SED fitting for a data set corresponding to the same source, whereas the two figures at the bottom, Fig c and Fig d, represent a clear case of bad SED fitting for a case of data set corresponding to different sources. These results validate the SED fitting algorithm defined.
We evaluated, however, the likelihoods of the hypotheses by integrating them over the entire parameter space.
We call Φ the parameter vector of the model H, which is simply a shorthand for writing (α, T, z). Similarly, we can consider (m, T, z) by applying a simple coordinate change between the brightness scale α, and the magnitude m in our model. These are the unknown model properties of the common object behind the observations. Generally the two Bayesian evidences which comprise the Bayes Factor here are expressed as follows. The numerator of B_{ph} is

$$p(D|H) = \int p(\Phi|H) \prod_i^N p_i(g_i|\Phi,H)\, d\Phi. \qquad (3.8)$$

where $p(\Phi|H)$ is the prior probability density distribution for the hypothesis H, and $p_i(g_i|\phi,H)$ is the likelihood of the data g_i given the model H. As $p(D|H)$ refers to the evidence for hypothesis H, this means that the measured flux in each band of the tuple corresponds to the same source, the prior probability will be the same.
With the likelihood function at hand, this can be numerically evaluated for a choice of any prior as we will see in the following sections.
We also considered the possibility of separate objects that correspond to the observations, hence the denominator of B_{ph} is a product of N integrals because separate sets of $\{\Phi_i\}$ parameters are required to accommodate this hypothesis:

$$p(D|K) = \prod_i^N \int p_i(\Phi_i|K)\, p_i(g_i|\Phi_i,K)\, d\Phi_i. \qquad (3.9)$$

where $p_i(\Phi_i|K)$ is the prior probability density distribution for the hypothesis K and the element i of the tuple and $p_i(g_i|\Phi_i,K)$ is the likelihood of the data given the model K. As $p(D|K)$ refers to the evidence for hypothesis K, meaning that all measured fluxes in each band of the tuple do not

correspond to the same source, the prior probability will be different for each element of the tuple. The photometric Bayes Factor can be obtained using the data and the model as described here; this Bayes Factor will determine which of the two hypotheses described above is the most plausible one in the probabilistic scenario that we present here. There is no prior preference for any of the hypotheses considered and this will make the photometric Bayes Factor the key element in the decision. We can resolve the overall astro-photometric cross-matching problem computing a combined Bayes Factor as the product of the astrometric and the photometric Bayes Factor as explained in [Budavári and Szalay, 2008]. A detailed definition of the data and the model is very important in order to represent the problem in a suitable manner. The prior probability distribution of the model parameters is often taken to be flat within a given parameter range. The SED model parameter brightness scale, α, can be expressed in terms of the apparent magnitude m for a range preliminary defined. The prior pdfs considered here are proper priors, therefore for the computation of the pdfs it is feasible to apply a change of the model parameters, using (T, z, m) instead of (T, z, α).
Starting with a simple case of 2-tuples with flux measured g_a and g_b and associated measurement errors σ_a and σ_b, we can write the evidences of the two hypotheses, H and K, as follows:

$$p(D|H) = \sum_{T,z,\alpha} p(T, z, \alpha|H)p(g_a|z, \alpha, T, H)p(g_b|z, \alpha, T, H)\Delta T\Delta z\Delta\alpha \tag{3.10}$$

where $p(T, z, \alpha|H)$ is the prior probability in terms of the parameters of the model and $p(g_a|z, \alpha, T, H)$ is the probability that a source in the bandwidth a with flux corresponding to a SED model of parameters α, T, z, which means, $\alpha f_a(T, z)$, has measured flux g_a. ΔT, Δz and $\Delta\alpha$ are the integration elements in each grid dimension. The prior probability is the same for both sources because hypothesis H implies that both sources of fluxes g_a and g_b correspond to the same galaxy and therefore their measurements will be modelled by the same SED template.

$$p(D|K) = \prod_{i=a,b} \sum_{T_i,z_i,\alpha_i} p(T_i, z_i, \alpha_i|K).p(g_i|z_i, \alpha_i, T_i, K)\Delta T_i\Delta z_i\Delta\alpha_i \tag{3.11}$$

where $p(T_i, z_i, \alpha_i|K)$ is the prior probability for each source of the tuple and $p(g_i|z_i, \alpha_i, T_i, K)$ is the probability that a source in the i band corresponding to $\alpha f_i(T, z)$ with specific parameters for that band, z_i, α_i, T_i has a measured flux g_i. For this hypothesis K, the elements of the tuple correspond to different galaxies and therefore the prior probabilities are different for each source involved. In this case we have two elements in the tuple for which the Bayesian inference yields the most plausible hypothesis for the two measurements belonging or not to the same galaxy.
Previously we only considered separate detections with their own measurements. When matching across surveys or instruments, one is often given multiple fluxes in each catalogue. A good example is matching the catalogue of the Sloan Digital Sky Survey (SDSS; [Shimasaku et al., 2001]) and the Galaxy Evolution Explorer (GALEX; [Martin et al., 2005]). The former provides *ugriz* magnitudes, while the latter has near and far ultraviolet fluxes (NUV and FUV for near- and far-ultraviolet). This can be considered as a two-way crossmatch problem with 5+2 measurements.
If, however, the GALEX band-merging algorithm were suspect due to its large astrometric uncertainties, one could consider this a three-way matching problem of SDSS, NUV and FUV catalogue with 5+1+1 fluxes so that the joint analysis could improve the quality of the associations.
Our statistical methodology can naturally deal with these scenarios and any other combination. The H hypothesis does not change, but the competing hypothesis K now has $\{G_k\}$ groups of measurements, hence its likelihood becomes

$$p(D|K) = \prod_{k}^{N_G} \int p_k(\Phi_k|K) \prod_{i\in G_k} p_i(g_i|\Phi_k, K)\, d\Phi_k, \tag{3.12}$$

where N_G is the number of groups. For N groups with an individual data point in each one, this becomes Equation 3.9. Also, formally, when all observations are in the same group, it becomes hypothesis H, but would not correspond to a matching problem.
This can be extended to a more complex problem, called the n-way approach where the tuple has n elements corresponding to n different pass-bands. In this case the Bayesian evidence for the hypothesis H and K can be expressed similarly as before:

$$p(D|H) = \sum_{T,z,\alpha} p(T,z,\alpha|H) \prod_{i=1}^{n} p_i(\vec{g_i}|z,\alpha,T,H)\Delta T \Delta z \Delta \alpha \tag{3.13}$$

where $p(T,z,\alpha|H)$ is the prior probability for each of the model parameters and $p(\vec{g_i}|z,\alpha,T,H)$ is the probability that a source in the bandwidth i with flux corresponding to a SED model of parameters α, T, z, which means, $\alpha f_i(T,z)$, has a measured flux g_i. The prior probability is the same for all sources of the tuple under consideration because hypothesis H implies that each source of flux g_i corresponds to the same galaxy and therefore their measurements will be modelled by the same SED template.

And

$$p(D|K) = \prod_{i}^{n} \sum_{T_l,z_l,\alpha_l} p_i(T_l,z_l,\alpha_l|K) \cdot p_i(\vec{g_i}|z_l,\alpha_l,T_l,K)\Delta T_l \Delta z_l \Delta \alpha_l \tag{3.14}$$

where n is the number of different instruments and l is the number of pass-bands of each instrument; this value l can be different for each instrument.
A natural boost of the photometric Bayes Factor is expected with the number of tuples involved in the computation. Therefore, the use of an $l-way$ approach when relevant will keep the framework in a more realistic scenario.
To assign the above probabilities, we need to be able to compare the data with the predictions for each model: the larger the mismatch, the lower the corresponding probability. Photometric accuracy can be modelled in terms of a Gaussian distribution. Thus, we can write the likelihood distribution as follows:

$$p_i(\vec{g_i}|z,\alpha,T,H) = \frac{1}{\sqrt{2\pi}\sigma_i} \cdot \exp\left(-\frac{(\vec{g_i} - \alpha\vec{f_i}(T,z))^2}{2\sigma_i^2}\right) \tag{3.15}$$

Then, the maximization problem of the likelihood can be obtained by minimizing the χ distribution. Therefore, the likelihood can be written as follows:

$$\mathcal{L} = p(\vec{g_i}|z,\alpha,T,H) \propto \exp\left(-\chi^2(\alpha,T,z)/2\right) \tag{3.16}$$

The brightness scale α will be expressed in terms of the apparent magnitude m for a preliminary-defined range. For the prior *pdf* we can implement the following change of variables:

$$\int p(m)dm = 1 = \int p(\alpha)d\alpha \tag{3.17}$$

This approach will allow the computation of the best fitted model and therefore the best fitting parameters at the same time as we compute the photometric Bayes Factor.

3.6.3 Influence of Uncertainty in the Photometric Bayes Factor

Uncertainty is always a topic of interest in the field of data analysis. Every piece of quantitative information in Science is linked to a certain degree of uncertainty which needs to be analysed in order to provide a realistic interpretation of the results. Uncertainties associated with the data (measured error), with the model (softening factor) and systematic uncertainties associated with the computational limit and efficiency of the algorithms proposed have been assessed to evaluate their impact in our problem.
Similar to the case of astrometric uncertainty, the calibration of the photometry of sources in a catalogue is obtained by comparing data values with the photometric values of the model SED templates from the adequate library. Today we can find for almost every survey an appropriate SED template library as a reference for the true values. It is important to understand the effects of uncertainty in the Bayes Factor result; this means understanding how uncertainty plays a role in the Bayesian inference. In this sense, the magnitude of uncertainty dm is approximately $dm \cong \frac{df}{f}$, f being the flux and df the uncertainty in terms of flux. In effect, let us describe the relationship between σ in terms of apparent magnitude and in terms of flux:

$$m = -2.5 \log f = -2.5 \log \frac{\ln f}{\ln 10} \tag{3.18}$$

$$|dm| \cong |-\frac{df}{f}| \cong \frac{\sigma_{flux}}{f} \tag{3.19}$$

$$\sigma_m \cong \|\frac{-2.5}{\ln 10}\| \frac{\sigma_{flux}}{f}$$

SED templates do not come with uncertainty; a possible way around is to soften the measurement errors to account for this. In addition to the above σ_{flux}, a softening factor accounting for the template uncertainty will be introduced in order to simulate error in the model. A typical softening is a few percent of the measured flux added in quadrature. Thus, in addition to the above df, a softening factor, or **Smooth Factor** η accounting for the template uncertainty, will be introduced in order to simulate error in the model. The value of uncertainty of flux on the model will be as follows:

$$\sigma^2_{\text{model}} = \sigma^2_{\text{flux}} + \eta^2_{\text{flux}} \tag{3.20}$$

It is very important to find an appropriate range of values for the Smooth Factor; if we have large Smooth Factor, the values of the Bayes Factor for the matched samples will shrink towards zero and there will not be much evidence coming from the data; if the value of the Smooth Factor is small but still realistic, the range of Bayes Factor values will be wider and therefore we will learn more useful information from the data. For our problem, the result of the implementation of various softening factors, η, let us say: $0.01, 0.03, 0.05, 0.1, 0.15$, was analysed to gain understanding of the influence of this factor in the overall uncertainty. It was concluded, as reported in [Marquez et al., 2014] that a Smooth Factor of 0.03 seems acceptable for the data used in this research.
Figure 3.11 shows the influence of the magnitude error on the Bayes factor. A checking of the code used in MATLAB for the generation of random numbers from a Gaussian distribution, along with an assessment of the performance of this algorithm, has been done. The code that we have used for random number generation is based on the *Mersenne Twister* algorithm and it is in line with the generation of random bits as explained in [Press et al., 2007]. The specific code used in MATLAB for the generation of random bits can be found in [Okumura, 2002] for a C++ version and in [Nishimura and Matsumoto, 2002] for a C version.
The algorithm described in [Matsumoto and Nishimura, 1998] is used in MATLAB, as well as in R, Python and PHP for example.
In order to test and validate our conclusions about the suitability of the MATLAB random generator

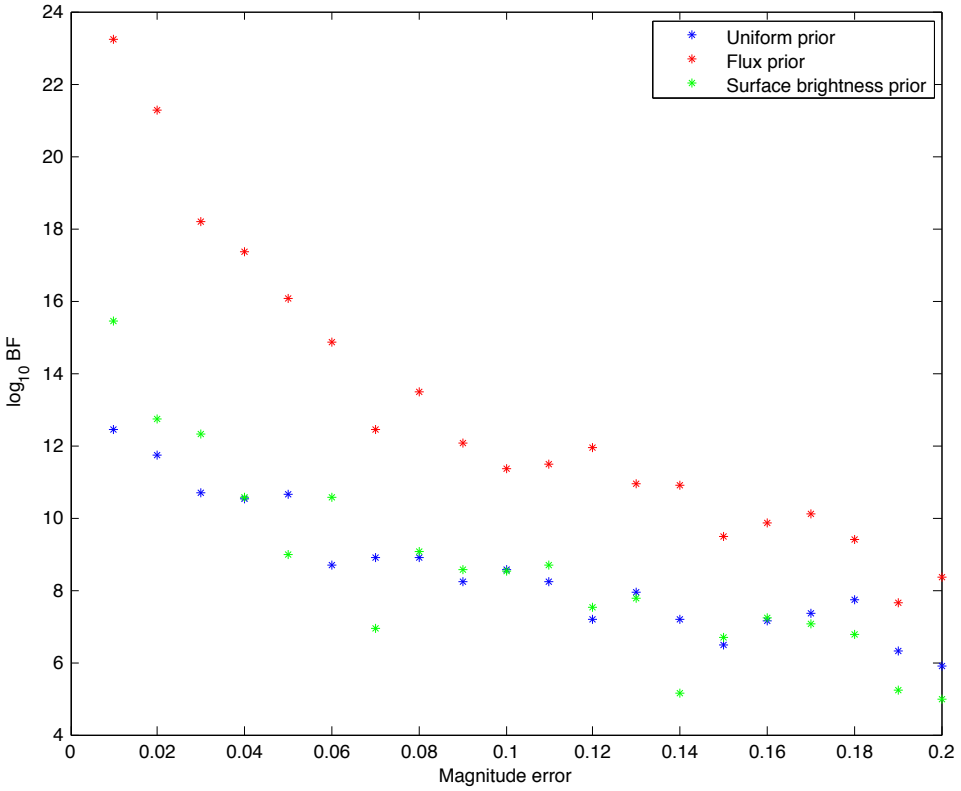

Figure 3.11: 8-way cross-matching (from Subaru B, V, R, G and IRAC bands $\log_{10} BF$ versus a range of magnitude error values simulating real data for the different priors considered in this research: uniform, flux and surface brightness (for more details on the priors, see Section 3.6.1). The Smooth Factor applied here is 0.03.

code, we have taken the following steps:

1 We ran the C++code of [Okumura, 2002] in a typical linux GCC environment and compared the results with the same execution in MATLAB (under the Mac OS X 10.6.8 environment in which every code in this research has been developed). From both environments, the histograms yield similar results.

2 We ran the MATLAB code 1000 and 10000 times to generate a specific *simulated perfect data* from a normal distribution and obtained the histograms below.

3 Finally, we repeated step 2 for a uniform distribution between 0 and 1.

Figure 3.12 contains the results of steps 2 and 3. From all this information we can conclude that the algorithm for generating random values used in MATLAB is valid for the purpose of our research.

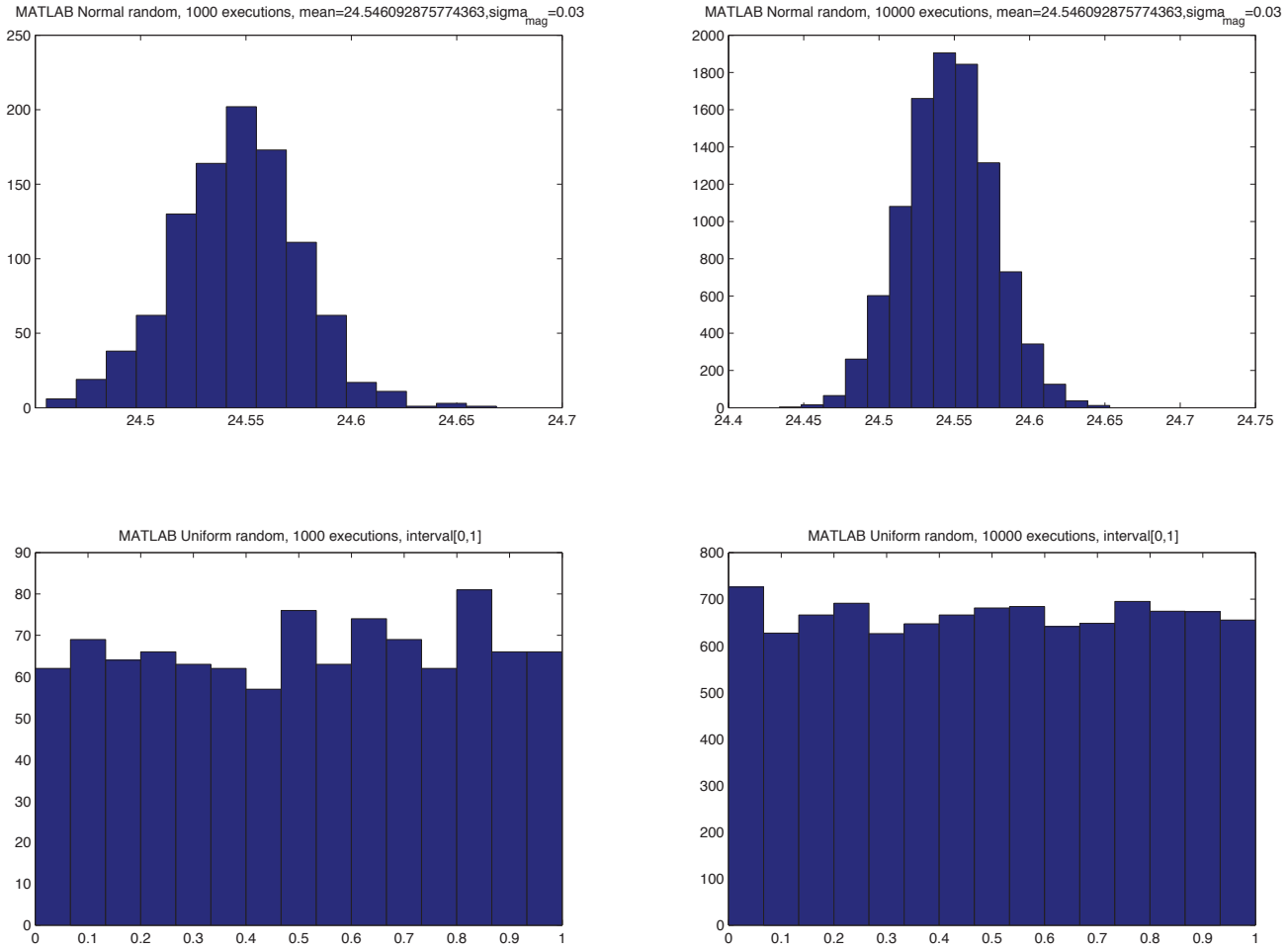

Figure 3.12: Results from the tests done in MATLAB random generator code.

A set of repeated executions were run and results assessed in order to determine if there are significant systematic errors. The outcome of those executions showed a random slight scatter

around the expected values; therefore we conclude that there are no significant systematic errors affecting the outcomes of the algorithms presented here.
A bench-marking of execution of this algorithm in C++ and MATLAB environments was carried out obtaining very similar results. Therefore we conclude that the algorithm used by MATLAB for the generation of random numbers following a Gaussian distribution is robust enough to base the computation of our research on the MATLAB environment.

3.6.4 Influence of Priors in the Photometric Bayes Factor

This section addresses the influence of the prior in the Bayes Factor. This comes in line with the fact that better knowledge about the problem to be solved will produce more accurate results in the computation of evidence for the model decision problem described above.
The term *subjective probability* is qualifying the discussions on Bayesian methods. The main point of concern tends to be the choice of the prior pdf and what to be done if this is not known. According to [Sivia and Skilling, 2006], this issue is often considered one of the main difficulties for implementing a Bayesian approach, and it is referred to as evidence of its subjective nature.
However, the wide concept of prior probability cannot be understood as something known; it is simply an assignment which reflects the relevant information available at the moment of expressing that probability. The requirement of consistency will ensure that the same knowledge will reach the same probability distribution function.
As stated in [Sivia and Skilling, 2006], the issue of *improper priors*, that is to say, a probability distribution function which cannot be normalised, is often quoted as a serious impediment for the general use of the Bayesian approach. In general, for parameter estimation, improper priors merely constitute a technical inconvenience rather than a serious difficulty. For model selection, however, the situation is quite different; the prior must be normalised properly because the evaluation of the probabilistic evidence entails an averaging of the likelihood function over it.
Assuming enough knowledge to describe the problem, various prior probability distribution functions are proposed and analysed in order to obtain some preliminary outcomes concerning their impact on the problem of hypotheses decision.

Uniform Priors

Considering no information a priori about the model proposed, we can assume a 3D uniform distribution, that is, separate uniform priors for each of the model parameters indicated above. Therefore the prior pdf can be expressed as follows:

$$p(T, z, m) = p(T)p(z|T)p(m|z, T) \tag{3.21}$$

The functional form of these pdfs is a constant value for each of the relevant parameters - T, z and m - in a defined range of values and such that the integral is one.
$p(T), p(z|T)$ and $p(m|z, T)$ follow uniform distributions for each parameter T, z, m identified in the model. Thus, we can develop the above equation for each element i of the model grid as follows:

$$p(T_i, z_i, m_i|H) = \frac{1}{\Delta(T_i)} \cdot \frac{1}{\Delta(z_i)} \cdot \frac{1}{\Delta(m_i)} \tag{3.22}$$

Then, the photometric Bayes Factor will be computed as the ratio between the two evidences, $p(D|H)$ and $p(D|K)$ based on the hypotheses H and K:

$$PhBF = \frac{p(D|H)}{p(D|K)} \tag{3.23}$$

where these two evidences will be obtained as the integral over all the range for all the parameters on which each hypothesis is dependent. Using a simple approach to numerical integration in terms of sums, we can write the following:

$$p(D|H) = \frac{1}{\Delta T}\sum_{T_p}\delta T_p \frac{1}{\Delta Z}\sum_{z_j}\delta z_j \frac{1}{\Delta m}\sum_{m_k}\delta m_k \cdot exp(-\sum_{i=1}^{n_{band}}\frac{(g_i-\alpha(m_k)f_i(z_j,T_p)^2}{2\sigma_i^2}) \tag{3.24}$$

where each of the first three sums refer to the prior probability in each of the parameters under consideration. This is summed over the range of each parameter in the model grid, defined respectively by the indices p, j and k, where δT_p, δz_j and δm_k are the step for the numerical integration of each parameter T, z and m; and $exp(-\sum_{i=1}^{n_{band}}\frac{(g_i-\alpha(m_k)f_i(z_j,T_p)^2}{2\sigma_i^2})$ is the likelihood of the flux g_i of the i band computed throughout all elements of the model grid, and is then multiplied in each band of the tuple.
Similarly, for hypothesis K we compute the evidence as follows:

$$p(D|K) = \prod_{i=1}^{n_{band}}\frac{1}{\Delta T_i}\sum_{T_p}\delta T_{pi}\frac{1}{\Delta Z_i}\sum_{z_j}\delta z_{ji}\frac{1}{\Delta m_i}\sum_{m_k}\delta m_{ki}\cdot\exp\left(-\frac{(g_i-\alpha(m_{ki})f_i(z_{ji},T_{pi}))^2}{2\sigma_i^2}\right) \tag{3.25}$$

where every term in this equation is under the product covering each member of tuple i because this hypothesis refers to the non cross-matching status of each source of the tuple. Note the double index in the sum terms (one refers to the tuple membership and the other one to the model grid membership).

Non-Uniform Prior: Surface Brightness

For a source of magnitude m which extends over an area of A $arcsec^2$, the surface brightness S is given by:

$$S = m + 2.5\log_{10}(A) \tag{3.26}$$

As mentioned in [Trotta, 2008], a new non-uniform prior has been proposed by using the concept of galaxy surface brightness. Its dimming is proportional to $(1+z)^{-4}$; therefore it can be used as a prior in our problem. The use of surface brightness as prior intends to break the redshift degeneracy as indicated in [Xia et al., 2009].
Finding dependencies between the parameters would allow us to simplify the above 3D integration to a lower order integral for the computation of the prior probability. These potential dependencies between the parameters of the model can be obtained with a better knowledge of the model itself. In this sense, one method of estimating distance ratios of galaxies, as described in [Schneider, 2014], uses the surface brightness fluctuations. It is based on the fact that the number of bright stars per area element in a galaxy fluctuates - purely by Poisson noise. If N stars are expected in an area element, relative fluctuations of $\frac{1}{\sqrt{N}}$ of the number of stars will occur. These are observed in fluctuations of the local surface brightness. To demonstrate that this effect can be used to estimate distances, we consider a solid angle $d\Omega$. The corresponding area element $dA = D^2 d\Omega$ has a quadratic dependency on the distance D of the galaxy. The larger the distance, the larger the number of stars N

in this solid angle, and the smaller the relative fluctuations of the surface brightness. By comparing the surface brightness fluctuation of different galaxies, one can then estimate relative distances.
In an expanding universe, there are two main effects which reduce the power detected from distant objects. First, the reduction of the rate at which photons are received. As each photon has to travel a little farther than the previous one, the rate is reduced. Second, the redshift, which reduces the energy of each photon observed. In addition to that, distant objects appear larger than they are in reality. This happens because the photons observed were emitted at a time in which the object was closer than at the time of measurement. Adding these effects together we conclude that the surface brightness should decrease with the fourth power of $(1 + z)$. All this is under the consideration of an expanding universe (flat geometry and uniform expansion over the range of redshifts observed). Based on this, we can propose the following dependency:

$$S \propto (1+z)^{-4} \tag{3.27}$$

Therefore for a source with magnitude m extending over an area A $arcsec^2$, we can propose the following dependency:

$$m \cong (1+z)^{-4} - 2.5\log_{10}(A) \tag{3.28}$$

Using the equivalence of equation 3.17, we can introduce the following prior:
$p(T, m, z) = p(m|z, T)p(z|T)p(T)$, where we have eliminated the influence of the parameter A by marginalizing as follows:

$$p(m|z) = \int_{A_{min}}^{A_{max}} p(m, A|z)dA \tag{3.29}$$

Therefore, we can write:

$$p(m|z) = C(A_{max} - A_{min})(1+z)^{-4} - \frac{2.5}{\ln 10}[(\ln A - A)_{A_{max}} - (\ln A - A)_{A_{min}}] \tag{3.30}$$

where A_{min} and A_{max} will be selected according to a reasonable range which generously covers the horizon of cases for our problem, and where C is a proportionality constant to ensure that the prior probability distribution is a proper prior distribution.
The functional form of this PDF belongs to the family of power functions, and again the value of the integral over the parameter range is one.
Thus, we can write the evidences using the above proposed joint prior as follows:

$$p(D|H) = \frac{1}{\Delta T}\sum_{T_p}\delta T_p \frac{1}{\Delta Z}\sum_{z_j}\delta z_j \sum_{m_k} p(m_k|z)\delta m_k \cdot exp(-\sum_{i=1}^{n_{band}} \frac{(g_i - \alpha(m_k)f_i(z_{bestfit}, T_p)^2}{2\sigma_i^2} \tag{3.31}$$

And for hypothesis K:

$$p(D|K) = \prod_{i=1}^{n_{band}} \frac{1}{\Delta T_i}\sum_{T_p}\delta T_{pi} \frac{1}{\Delta Z_i}\sum_{z_j}\delta z_{ji} \sum_{m_k} p(m_{ki}|z)\delta m_{ki} \cdot \exp(-\frac{(g_i - \alpha(m_{ki})f_i(z_{bestfit}, T_{pi})^2}{2\sigma_i^2}) \tag{3.32}$$

A step further would be to consider the 2D prior: $p(T, m(z)) = p(T)p(m(z)|T)$ by marginalizing over an area A and the redshift z:

$$p(m) = \int_{z_{min}}^{z_{max}} \int_{A_{min}}^{A_{max}} p(m, A, z)dAdz \tag{3.33}$$

Developing the above integral analytically, we obtain:

$$p(m) = C(A_{max} - A_{min})\Big(-\frac{1}{3}\Big)\Big(\frac{1}{(1+z)^3_{z_{max}}} - \frac{1}{(1+z)^3_{z_{min}}}\Big) - \\ -\frac{2.5}{\ln 10}[(\ln A \cdot A - A)_{A_{max}} - (\ln A \cdot A - A)_{A_{min}}](z_{max} - z_{min}) \tag{3.34}$$

where again the values for $A_{min}, A_{max}, z_{min}, z_{max}$ will be selected according to an appropriate range for the problem under consideration and where C is the proportionality constant to ensure that the prior distribution is a proper prior.
Here it is expected that the computational effort will considerably decrease if we keep as the value of α the one corresponding to the minimum χ, i.e. best fitting, for the parameter z of the grid introduced.
Then we can write the evidences using the joint prior proposed above, as follows:

$$p(D|H) = \frac{1}{\Delta T}\sum_{T_p}\delta T_p \sum_{m_k} p(m_k)\delta m_k \cdot exp(-\sum_{i=1}^{n_{band}}\frac{(g_i - \alpha(m_k)f_i(z_{bestfit}, T_p))^2}{2\sigma_i^2} \tag{3.35}$$

$$p(D|K) = \prod_{i=1}^{n_{band}}\frac{1}{\Delta T_i}\sum_{T_p}\delta T_{pi}\sum_{m_k}\delta m_{ki} \cdot exp(-\frac{(g_i - \alpha(m_k)f_i(z_{bestfit}, T_p))^2}{2\sigma_i^2} \tag{3.36}$$

Where $z_{bestfit}$ indicates the value of the redshift with which the best SED fit is achieved.
As we will see in the results presented in the next section, the use of a surface brightness prior for the data set used here slightly improves the computational cost, but it does not greatly improve the classification compared to the use of uniform prior.

Prior Based on Data Fluxes: Flux Prior

This paragraph describes the prior proposed in [Fadely et al., 2012] adapted to the case of our problem of multi-band galaxies cross-matching. In this case we will use a sub-set of data as part of the prior information. The size of this subset represents the range of the entire data set in order to ensure that the results are not overly sensitive to the sample size of the data. For the priors on template type T and on redshift z we consider uniform distributions. Similarly as done in [Fadely et al., 2012], we will marginalize over the fitting for the scale brightness parameter within a range of $\pm 3\sigma$. Thus, the prior probability density distribution for the scale brightness α will be approximated by a Gaussian distribution where the mean and deviation are obtained by computing the weighted average and variance for each model SED fitting with the data subset, those being the hyper-parameters β which parametrise the scale brightness prior distribution. These hyper-parameters, chosen such that their sum is equal to unity, are in our case the mean β_{wmean} and variance β^2_{wvar} of a log normal distribution on the brightness scale factor α. It is therefore the simultaneous inference of the hypotheses H, K and of the hyper-parameters β what makes this approach hierarchical.
The expected benefit of this approach compared with others, presented in the above paragraphs of this section, is that a subset of the data provides information about the prior probabilities for each individual source.
We reduce here the 3D Uniform prior defined above, $p(T, z, \alpha)$ to a 2D Uniform prior in type T and redshift z, $p(T, z) = p(T)p(z)$. To achieve that, we marginalize the likelihood over the scale brigthness parameter α considering the uncertainties of the scale brightness for the best fitting.

Then the likelihood can be written:

$$p(g_i|z,\beta,T,H) = \int_{-3\sigma_\alpha}^{3\sigma_\alpha} p(g_i|T,z,\alpha,\beta,H)p(\alpha|T,z,\beta,H)d\alpha, \tag{3.37}$$

where α is the overall scale brightness factor and β represents the hyper-parameters used when marginalizing over the scale brightness fitting uncertainty. This will result in:

$$p(\alpha|T,z,\beta,H) \cong C\exp\left\{-\frac{1}{2}\frac{(\ln\alpha_{val}-\beta_{wmean})^2}{\beta_{wvar}^2}\right\}\cdot\frac{1}{\alpha_{val}}, \tag{3.38}$$

where a log-normal distribution was chosen to represent the probability density distribution of the scale brightness factor. The proportionality constant C is chosen such that the prior probability of α is a proper prior. Therefore, α_{val} is the result of marginalizing the brightness scale considering a data subset and the model grid.
This implementation significantly improves the computational cost if we compare it with the implementations described in the two previous sections.
Let $\vec{d}$ be the subset of data values for this problem, i.e., the tuples N_d of N_b bands, and let $\vec{e}$ be the uncertainty in the data. The size of $\vec{d}$ is $N_d x N_b$ where N_d is the number of tuples which comprises the data subset and N_b is the number of bands in each tuple for the multi-band problem under consideration. Then, for each SED template SED_i where i denotes a specific type and redshift, the two hyper-parameters for the prior of scale brightness will be, as indicated above, the weighted mean, β^i_{wmean} and the weighted variance β^i_{wvar} of the fitting of all data subset against the SED_i template.
For the sake of readability, the super-index i will be suppressed; it being taken as understood that the following expressions are referred to a specific SED template SED_i. Thus size of β_i for each specific SED template, SED_i, is $N_d \cdot N_b$.
We define the following coefficients in the β hyper-parameters computation.

$$\ln\mu_w = \ln\frac{\sum_{j=1}^{N_b}\frac{\vec{m_j}\cdot\vec{m_j}}{\vec{e_j}\cdot\vec{e_j}}}{\sum_{j=1}^{N_b}\frac{\vec{d_j}\cdot\vec{m_j}}{\vec{e_j}\cdot\vec{e_j}}} \tag{3.39}$$

for the coefficient used to compute the weighted mean, where $\vec{m_j}$, $\vec{d_j}$ and $\vec{e_j}$ are respectively the corresponding values of the model, data and uncertainty tuples.

$$\sigma_w = \sqrt{\frac{1}{\sum_{j=1}^{N_b}\frac{\vec{m_j}\cdot\vec{m_j}}{\vec{e_j}\cdot\vec{e_j}}}} \tag{3.40}$$

for the coefficient used to compute the weighted deviation.
We define the weight as:

$$W = \frac{1}{\sigma_w^2} \tag{3.41}$$

We also define the following intermediate factors for the computation of the weighted variance:

$$V1 = \sum_{j=1}^{N_d} W_j \tag{3.42}$$

$$V2 = \sum_{j=1}^{N_d} W_j^2 \tag{3.43}$$

$$V = \frac{V1^2}{V1^2 - V2} \tag{3.44}$$

where V is the correction applied to the weighted variance for the effective number of points. Then we can write the scale brightness hyper-parameters for each SED_i template as follows:

$$\beta_{wmean} = \frac{\sum_1^{N_d} \ln \mu_w \cdot W}{V1} \tag{3.45}$$

$$\beta_{wvar} = V \cdot \sum_1^{N_d} W \cdot (\ln \mu_w - \beta_{wmean})^2 \tag{3.46}$$

We will use uniform distributions for the prior probability distributions for T and z, where the hyper-parameters are the weights which fulfil the proper prior condition.
Expanding the above expressions following the hierarchical Bayesian approach and marginalizing over the scale brightness, we obtain the following equation for the evidence of hypothesis H:

$$p(D|H) = \int_T p(T|H)dT \int_z p(Z|H)dZ \int_\alpha max(p(\vec{g_i}|\alpha, Z, T, \beta, H))p(\alpha|T, Z, \beta, H)d\alpha \tag{3.47}$$

where the prior of T and of z have been considered independent, and the likelihood has been approximated for this hypothesis H, by its maximum i value among all bands N_{band} in terms of the minimum χ value. It has also been marginalized over the range of possible scale brightness values parametrised with their uncertainties in their fitting among all the data sets.
Using a simple numerical integral approximation, we can develop the above equation as follows:

$$p(D|H) \cong \sum_t \frac{1}{\Delta T}\delta t \sum_z \frac{1}{\Delta Z}\delta z \sum_\alpha \exp(-\frac{1}{2}\chi^2_{min}) \exp(-\frac{(\alpha_{val} - \beta wmean)^2}{2\beta wvar}) \frac{1}{\alpha_{val}}\delta\alpha \tag{3.48}$$

For the hypothesis K, the likelihood accounts for the non cross-match; therefore the approximation of its maximum value among all the bands cannot be used here. In this case, the evidence can be written as follows:

$$p(D|K) = \prod_{i=1}^{N_{band}} \int_{T_i} p(T_i|K)dT_i \int_{Z_i} p(Z_i|K)dZ_i \int_{\alpha_i} p(\vec{g_i}|\alpha_i, Z_i, T_i, \beta, K)p(\alpha_i|T_i, Z_i, \beta, K)d\alpha_i \tag{3.49}$$

Using a numerical integral approximation as for the evidence of hypothesis H we can write the following for the evidence of hypothesis K:

$$p(D|K) \cong \prod_{i=1}^{N_{band}} \sum_{t_i} \frac{1}{\Delta T_i}\delta t_i \sum_{z_i} \frac{1}{\Delta z_i}\delta z_i \sum_{\alpha_i} \exp(-\frac{1}{2}\chi^2_i) \exp(-\frac{(\alpha_{val_i} - \beta_{wmean})^2}{2\beta_{wvar}}) \frac{1}{\alpha_{val}}\delta\alpha_i \tag{3.50}$$

3.7 Source Contour Extraction in Astronomical Images

The measurement of the surface brightness of extended astronomical sources becomes especially challenging when we are dealing with faint sources, noisy images, crowded and confused areas with blended sources, etc. The traditional methodology consists of computing the pixel brightness above a detection threshold enclosed in a geometry based on the object light distribution (e.g. Kron ellipses). The astronomical images and associated catalogues from real surveys capture numerous faint sources, crowded areas with overlapped sources, etc. These catalogues contain quality flags to determine the reliability level of the surface brightness results. Photometry in astronomy is performed by computing the brightness of the sources in the astronomical survey images. In a general mathematical sense, image formation can be perceived as a process which transforms an input distribution into an output distribution. In this sense, the calibration of digital images takes on a considerable relevance in bringing accurate information to any image analysis and processing pipeline. Current astronomical surveys normally deal with multi-wavelength high precision images, assumed to be sufficiently accurate (with a level of uncertainty as given in the smooth factor earlier in this chapter).
A functional representation of the capturing process - the point spread function (PSF), p, describes the way that the information on the object function is spread as a result of recording the data. This is strongly dependent on the characteristic of the imaging instrument and the context in which it is used, and it is a deterministic function. Additionally, the image will contain additive noise n, which is a non deterministic function which can, at best, only be described in terms of some statistical noise distribution. We can think of noise as a stochastic function which is a consequence of all the unwanted disturbances that occur during the recording of the image data. All these factors are essentially combined as follows to form an image:
$s = p * o + n$, where $*$ represents the convolution operator.
As indicated in [Capak et al., 2007], the main barrier to effective image processing in general is noise and confusion in crowded areas. In this sense, improved calibration processes, like the one described in [Capak et al., 2007], along with technological advancements in the sensitivity capabilities of telescope instruments can contribute to improving the quality of the astronomical measurements.
As explained in [Osher and Fedkiw, 2006], a consistent PSF within each band and between bands is essential for high quality photometry and ideally all bands would have an identical PSF. However, due to the non-Gaussian portion of most PSFs, it is extremely difficult to consider the diverse data of surveys, such as COSMOS, which can achieve a homogeneous PSF.
The computation of surface brightness normally used in galaxies and other extended sources involves a first step of identifying the coordinates of the sources, and a second step of contouring the source and summing up pixel brightness within that contour. For the contour definition there are traditionally various options which produce subtle alternative results. Approximated elliptical contours based on the source light distributions seem to be generally used. Details can be found in [Bertin and Arnouts, 1996].

3.7.1 Improved Active Contour

In this section we introduce and benchmark two different approaches for the problem of defining the contour of galaxies in FITS astronomical images with crowded populations of galaxies including faint galaxies, such as the ones taken from the COSMOS survey.
Basically, the first approach defines the contour based on gradient of brightness in the image, while the second approach does not intend to explicitly determine the edges of galaxies but to group areas of the image with similar brightness levels.

First Approach: Active Contour Algorithm

In general an edge in an image can be considered as a discontinuity or gradient within the image. Edge detection is basically a method of segmenting an image into regions based on discontinuity. As indicated in [Chan and Vese, 1999], low-level feature detection processes are effective up to a point, but cannot be expected to retrieve entire geometric structures. In this sense active contour modelling with snakes techniques have constituted a fundamentally new approach. The basic idea in active contour modelling is to evolve a curve u_0, subject to constraints, from a given image in order to detect objects in that image. Ideally we begin with a curve around the object to be detected, and the curve moves normal to itself and stops at the boundary of the object. The classical snakes model involves an edge detector, which depends on the gradient of the image, to stop the evolving curve at the boundary. In a sequence of papers beginning with [Chan and Vese, 1999], a different active contour model without such stopping function is proposed. In this model, the stopping term is based on the Mumford-Shah segmentation technique, as described in [Chan and Vese, 2001]. The main advantages of this model against the classical snakes ones are: it can detect contours both with and without gradient; interior contours are automatically detected, and the initial curve can be anywhere in the image.
The images we are working with - astronomical images- do not in general have clearly defined edges. For the sake of the self-standing readability of this section, we present here a brief literal summary of the description of the model and the algorithm published in [Chan and Vese, 1999].
The method proposed here is based on the minimization of energy-based segmentation on the active contour model. For the sake of stating the principles of the methodology applied, it is assumed that image u_0 is formed by two regions of approximately piece-wise constant intensities of distinct values, u_0^i and u_0^0. The object to be detected is represented by the region with the value u_0^i and its boundary, C_0 is the result of minimizing the following fitting term:

$$F_1(C) + F_2(C) = \int_{A_I} |u_0(x,y) - c_1|^2 dxdy + \int_{A_O} |u_0(x,y) - c_2|^2 dxdy \tag{3.51}$$

where x and y are the coordinates in the image under consideration; $u_0(x,y)$ is the value of the brightness intensity of the image for the position (x,y), C is the evolving contour, and the constants c_1 and c_2, depending on C, are the average of u_0 inside C and respectively outside C. With this it is obvious to see that C_0, the boundary of the object, is the minimizer of the fitting term. A_I and A_O represents respectively the area inside and outside C.
In [Chan and Vese, 1999], the following energy function $F(c_1, c_2, C)$ is introduced and consists of the above fitting term and some regularizing terms, such as the length of curve C, and the area of the regions inside C.

$$\begin{aligned} F(c_1, c_2, C) = \mu \cdot Length(C) + \nu \cdot A_I + \\ \lambda_1 \int_{A_I)} |u_0(x,y) - c_1|^2 dxdy + \\ \lambda_2 \int_{A_O} |u_0(x,y) - c_2|^2 dxdy \end{aligned} \tag{3.52}$$

where μ is the parameter that minimizes the contour length, ν minimizes the curvature of the contour and by such impact in its smoothness. Then λ_1 and λ_2 are the parameters in charge of minimizing the contour distance from the object edges from inside and outside of the object respectively.
Following the level set formulation of the model presented in [Chan and Vese, 1999], the variable

curve C is represented by the zero level set of a Lipschitz function $\phi : \omega \to \mathbf{R}$, such that:

$$\begin{aligned} C = \partial\omega &= \{(x,y) \in \Omega : \phi(x,y) = 0\} \\ inside(C) = \omega &= \{(x,y) \in \Omega : \phi(x,y) > 0\} \\ outside(C) = \omega &= \{(x,y) \in \Omega \backslash \neg\omega : \phi(x,y) < 0\} \end{aligned} \tag{3.53}$$

The level set formulation of the variational active contour presented in [Chan and Vese, 1999] replaces C by ϕ and considers the following:

$$H(\phi) = \begin{cases} 1, & \text{if } \phi \geq 0 \\ 0, & \text{if } \phi < 0 \end{cases}$$

And

$\delta_0(\phi) = \frac{d}{d\phi} H(\phi)$

In order to compute the associated Euler-Lagrange equation for the unknown function ϕ, [Chan and Vese, 1999] considers slightly regularized versions of the functions H and δ_0, denoted by H_ϵ and δ_ϵ as $\epsilon \to 0$. Therefore the associated regularized functional can be defined as follows:

$$\begin{aligned} F_\epsilon(c_1, c_2, \phi) = \mu \int_\Omega \delta_\epsilon(\phi(x,y)) |\nabla\phi(x,y)| dxdy + \nu \int_\Omega H_\epsilon(\phi(x,y)) dxdy \\ + \lambda_1 \int_\Omega |u_0(x,y) - c_1|^2 H_\epsilon(\phi(x,y)) dxdy \\ + \lambda_2 \int_\Omega |u_0(x,y) - c_2|^2 (1 - H_\epsilon(\phi(x,y))) dxdy \end{aligned} \tag{3.54}$$

According to [Chan and Vese, 1999], keeping c_1 and c_2 fixed, and minimizing F_ϵ with respect to ϕ, the associated Euler-Lagrange equation for ϕ can be deduced. Parametrizing the descent direction by an artificial time $t \geq 0$, the equation in $\phi(t,x,y)$ (with $\phi(0,x,y) = \phi_0(x,y)$ defining the initial contour) is:

$$\begin{aligned} \frac{\partial\phi}{\partial t} = \delta_\epsilon(\phi) \left[\mu\nabla \cdot \left\{ \left(\frac{\nabla\phi}{|\nabla\phi|} \right) \right\} - \nu - \lambda 1 (u_0 - c_1)^2 + \lambda_2 (u_0 - c_2)^2 \right] = 0; \quad in \quad (0,\infty) \times \Omega \\ \phi(0,x,y) = \phi_0(x,y); \quad in \quad \Omega \\ \frac{\delta_\epsilon(\phi)}{|\nabla(\phi)|} \cdot \frac{\partial\phi}{\partial\vec{n}} = 0; \quad on \quad \partial\Omega \end{aligned} \tag{3.55}$$

Where $\vec{n}$ denotes the exterior normal to the boundary $\partial\Omega$, and $\partial\phi/|\partial\vec{n}$ denotes the normal derivative of ϕ at the boundary.
The regularized versions considered for H and δ are the following:

$$\begin{aligned} H_\epsilon(z) &= \frac{1}{2} \left(1 + \frac{2}{\pi} arctan \left(\frac{z}{\epsilon} \right) \right) \\ \delta_\epsilon(z) &= \frac{\partial H}{\partial z} = \frac{\epsilon}{\pi} \frac{1}{\epsilon^2 + z^2} \end{aligned} \tag{3.56}$$

The numerical approximation of this model is based on the finite differences method with h the space step and Δt the time step.

Refer to [Chan and Vese, 1999] for details of the numerical approximation to this model and the corresponding numerical computation of the associated Euler-Lagrange equation which describes the evolution of the contour.
The main steps of the algorithm proposed in [Chan and Vese, 1999] are the following:

1 Initialise ϕ^0 by $\phi_{0,n}$

2 Compute $c_1(\phi^n)$ and $c_2(\phi^n)$

3 Solve the PDE in ϕ to obtain ϕ^{n+1}

4 Reinitialise ϕ locally to the signed distance function to the curve (this step is optional)

5 Check whether the solution is stationary. If not, $n = n + 1$ and repeat

Until here the literal summary from [Chan and Vese, 1999].
The experiments presented in [Chan and Vese, 1999] were conducted with various number of iterations of the algorithm. Moreover, the length parameter μ, which has a scale role, was chosen according to the type of image in terms of noise, number and size of objects in the image. In this sense, if the objective is to detect all or as many objects as possible and of any size, then μ should be small. If the objective is to detect only larger objects (for example groups of objects), then μ should be larger. However, there is not a clear and quantitative understanding of the right values of this length factor for each image. The images used in this research will contain high populations of extended diffuse sources, and the perspective of performing several manual steps of trial and error to adjust the right value of μ seems inefficient for the objectives proposed here.
In order to improve the method described above for the challenging astronomical images with which we work, we proposed a new dynamic iterative contour stabilization condition based on a set of conditions addressing the difference in gradient of brightness and the number of foreground pixels versus background pixels from two consecutive iterations of the algorithm described above. We also keep a maximum number of iterations to avoid or minimize propagation of the contour through various contiguous extended blended sources in noisy images; we also introduced a dynamic iterative change in parameters λ_1 and λ_2 and for specific cases in ϵ, in order to optimise the speed of convergence of the algorithm and to define a contour which is not very much dependent of the characteristics of the specific image. In this sense, the weight of parameter μ in the performance of the algorithm is better balanced by the weight of the other factors, λ_1 and λ_2.
Let $\delta_i(\phi_i(p_j), C_i(p_0))$ be the delta value of brightness between each point p_j of the contour ϕ at the iteration i and the brightness of the image at the center p_0 of the source of which we are computing the contour C_i. Let $\overline{\delta_i}$ be the average of these values for each point of the contour j and for each iteration i. Finally, let β_i be the ratio of number of pixels of the contour C_i which belong to the background of the image versus the number of pixels of the contour C_i belonging to the foreground, for the iteration i.
Then, the algorithm implemented is as follows:

1 initialise ϕ^0 by ϕ_0,$n = 0$

2 Compute $c_1(\phi^n)$ and $c_2(\phi^n)$

3 Solve the PDE in ϕ to obtain ϕ^{n+1}

4 if $\overline{\delta_{i-1}} > \overline{\delta_i}$ finish algorithm and yield results.

Otherwise, if $\beta_{i-1} > \beta_i, \quad \epsilon = \epsilon/10$, to refine the contouring in the last steps of the algorithm.

5 Check whether the solution is stationary. If not, $n = n + 1$ and repeat

As we can see, we introduce in step [4] our modification to the algorithm presented in [Chan and Vese, 1999]. In this manner we introduce a refinement, based on statistics from brightness gradient, in the iterative resolution of the PDE. In this research the consideration is focused on the algorithm, therefore we assume that the instrument capabilities always achieve the best performance. With this in mind, we propose here an algorithm capable of dynamically finding the best set of conditions to achieve an optimum stable contour. In the following sections we present the main outcomes of its implementation, including an assessment on the robustness of the contour determination against various images with different noise ratios. Finally, we compare the results obtained with traditional tools such as SExtractor, with special focus on those cases in which the catalogue quality flag mask indicates poor quality.

Second Approach: Contour function of MATLAB

The second alternative approach investigated here consists of implementing the MATLAB contour function in one step for the complete image as described in the section above. Here the contours are simultaneously obtained for each of the galaxies in the image and the accuracy of the contours directly depends on one parameter related to the number of contour levels considered. This number of levels is connected with the different levels of brightness in the image.
This second approach consists of considering the images which are normally very crowded as a grid of data at different brightness levels. Generally this approach considers the contour as the intersection of three-dimensional surfaces with horizontal planes, yielding the contours as the lines at the same brightness level Z as result of this intersection. As described in [Hasanah et al., 2013], the contour plot consists of:

- Vertical axis: independent variable (y value)
- Horizontal axis: independent variable (x value)
- Lines: iso-response values (values with the same z value)

One important drawback of this approach is the lack of perception of discontinuities, which is crucial for the nature of the images we consider in this research.
.

Discussion

We assumed that the ratio S/N of the images would allow us to differentiate the background from the foreground properly in the majority of the cases, specially for bright sources.
When comparing both approaches presented above, it is important to note that both solutions retain different conceptual approaches yielding a difference in computational cost. The first approach considers each source in the image as an individual object, whereas the second approach considers all objects of the image as a unique object with different brightness levels enclosed in different separated contours.
In terms of computational cost, the second approach is considerably lighter than the first approach. However, and despite the promising preliminary results for the second approach, when extending that solution to a statistical representative number of cases, the first approach retains better adaptability to individual sources and, by that, to each contour. This becomes very important when dealing with overcrowded areas of sky where the noise of the image is an important factor and thus the individual approach to each source appears to be decisive to yield better overall results. We included here this brief analysis as a justification for deciding the first approach over the second one. A full detailed assessment on this will be further elaborated in Chapters 5 and 6. In order to improve the performance of our two approaches in areas of confusion, a combination of our algorithm with the use of Voronoi tessellation opens the door to possible further optimisation within this complex

problem.

Faint Galaxies Improving accuracy in the computation of the magnitude of faint galaxies is very challenging because we approach the boundaries of instrument limitations with these measurements and thus the signal to noise ratio is low. Some direct consequences from this are the following: the active contour algorithms encounter more difficulties in the finalization conditions and systematic errors are more frequent.
Both approaches presented here have taken this into account by dynamically adapting the parameters involved in the finalization condition of the algorithm. The only difference between both approaches is that, for the first one, the parameter adaptation takes place dynamically through the iterative process of the algorithm, whereas for the second approach, the selection of contour levels is static and does not take place in the contouring algorithm. In this sense, for the first approach, a noisy image will make the contour evolve more cautiously than a for less noisy image. This adaptability on the evolution of the contour is achieved by refining parameters λ_1, λ_2 and ϵ in the algorithm iterations. In the second approach, a good selection of the contour level parameter before running the algorithm will enable us to obtain optimum contours of the galaxies.
In view of the results obtained and described in Chapter 4, the robustness against noise of the first approach (modified active contour) presented here is promising. The same artificial image was recreated with and without noise and both algorithms were executed in both artificial images. The results obtained were very similar between both approaches and for both images.

Case of Overlapped Sources Another scenario where the performance of image processing may be compromised is when dealing with crowded areas and blended sources. In these cases, the capability of identifying isolated sources from overlapped ones is very important for obtaining more accurate surface brightness values.
Regarding overlapped sources, both approaches face difficulties. It is important to note here that, considering the traditional approach taken with Kron ellipses, it is obvious that without a human manual refinement step, the brightness of the pixels belonging to more than one Kron ellipse are automatically counted more than once for the relevant overlapped contours.
We propose here for a further study the combined use of a Voronoi tessellation with a pixel-based search algorithm to refine the frontier between two or more overlapped sources. This problem could itself be however the topic of another research and therefore the implementation of a solution for the overlapped problem will not be considered in the scope of this research beyond the approach proposed.

3.8 Labelling Isolated Galaxies

This section presents the approach followed in the classification of isolated versus non-isolated galaxies. Following a similar path as for the other contributions of this research, the detailed implementation of this approach and its validation is also part of this section.

3.8.1 Voronoi Tessellation in 2D Images

In this research, we create a Voronoi diagram by taking pairs of points that are close together and drawing a line that is equidistant between them and perpendicular to the line connecting them. That is to say, all points on the lines in the diagram are equidistant to the nearest two (or more) source points. These points, also named seeds or generators correspond to the sources coordinates,

and they are known beforehand. Then, for each seed, there is a corresponding region consisting of all points closer to that seed than to any other. These regions are called **Voronoi cells**.
The Delaunay triangulation is the mathematical duality of the Voronoi diagram. Let P be a set of points in a plane. Then, the Delaunay triangulation $DT(P)$ determines a triangulation such that none of these P points is inside the circumcircle of any triangle in $DT(P)$.

3.8.2 Rule-based system

In the first stage of our analysis pipeline, the catalogue extraction tool SExtractor described in [Bertin and Arnouts, 1996] is applied to each image separately. The catalogue obtained (including astrometric and photometric information) is used as the basis for a 2D Voronoi (Delaunay) tessellation of the images which defines a polygon in the corresponding coordinates (e.g. celestial, pixel) for each source.
MATLAB includes a few functions for obtaining Voronoi diagrams on a Euclidean space. We use the function $[V, C] = Voronoin(X)$ where V are the vertices of the Voronoi cells C.
The starting point is set to infinity. Therefore, depending on our problem, we can find that the first row of V is a point at infinity, which indicates that this Voronoi cell is unbounded.
The MATLAB 2D Voronoi tessellation of the astronomical images mentioned above provides us with a preliminary categorization of sources into the candidate categories of isolated source and partially or totally contaminated by neighbouring sources. As described in [Marquez, 2012], A source is labelled as the candidate for isolated source if it is fully contained in its Voronoi cell and none of the sources from the Voronoi cells surrounding the source under consideration is contaminating it. Initially, the source extension can be defined by its Kron ellipse, [Bertin and Arnouts, 1996] although subsequent refinements can be applied with more refined contours (active contours for example). This labelling procedure only considers information from one single image. The definition can be extended by defining isolated source as one which is i) isolated in the lowest resolution image; ii) only has one counterpart in the projection of its Voronoi cell in all other images and, iii) each of these counterparts is also isolated in the sense defined above. If any of the surrounding sources extends its contour outside its Voronoi cell and occupies part of the Voronoi cell of the source under consideration, then we change its label to potentially contaminated. Finally, if the contour of the source under analysis is not fully contained in its Voronoi cell, that source may be a contaminating one. Considering a collection of images corresponding to a multi-wavelength survey, a source will be isolated if it fulfils the conditions described above for that source in all the wavelengths of the survey under consideration.
Following this approach, we have implemented in our pipeline the step of labelling the sources according to their potential condition of being isolated, partially contaminated or contaminated.

Figure 3.13 shows an example of the result of the implementation of this preliminary labelling process on the Hubble Deep Field image taken by the IRAC instrument on channel $3.6\mu m$.
A simple improvement of this approach consists of taking the source morphology into account in the determination of the isolation cell instead of the SExtractor Kron ellipses yielded by the catalogues.
As concluded in [Marquez, 2012], the result from the previous steps will produce a set of two-dimensional vectors, $\vec{x_{ij}}$ which represent the celestial coordinates of the source j in catalogue i together with the preliminary labelling described in the previous paragraphs. From this set of vectors, we aim to construct reliable SEDs by cross-matching them, taking into account the astrometric information, the photometric information and the instrument sensitivities.

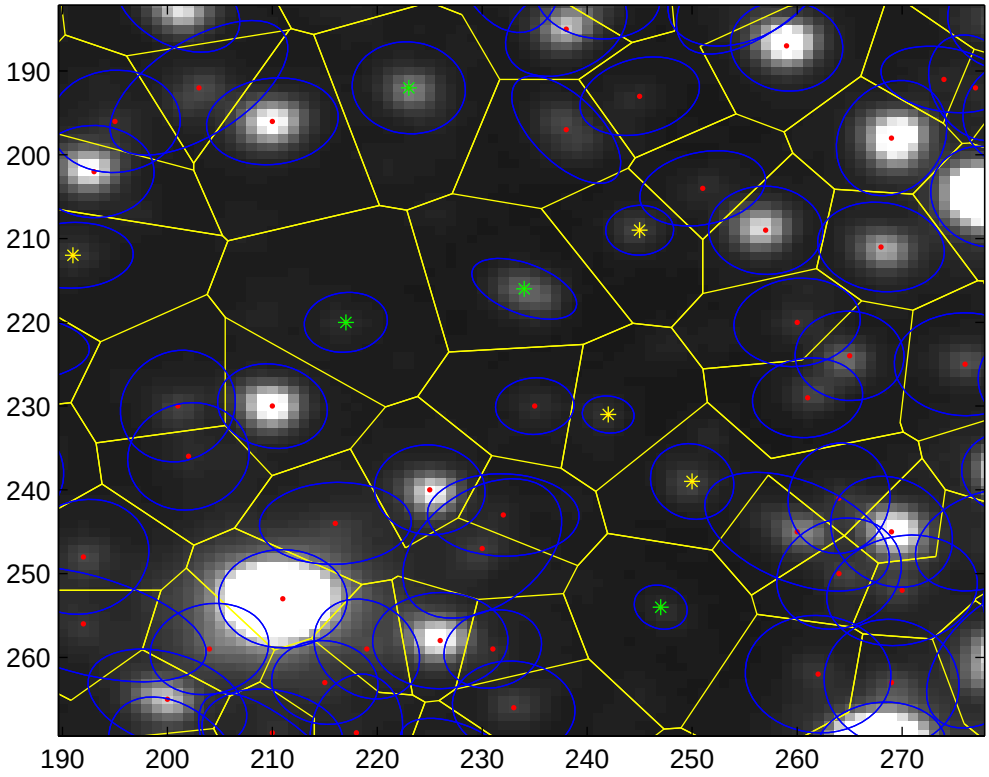

Figure 3.13: Examples of isolated and partially contaminated sources in the Hubble deep field image of IRAC instrument on channel $3.6\mu m$.

Chapter 4

Validation of the GCES

4.1 Introduction

Any new software prototype requires a validation of its intended functionality and verification of the defined requirements in a controlled environment. This environment is obtained by means of creating synthetic data, to validate the proposal.
One of the main goals when validating our expert system is to reach an adequate level of confidence of the suitability for its use with real data sets. In this chapter we explain the methodology used for the validation of our expert system (GCES) with synthetic data. We present the results on representative synthetic data sets and based on it, we perform a statistical study of these results and provide an assessment of the contributions and limitations of this prototype.
The main purpose of the synthetic data is to validate our GCES system, whereas the main role of the real data is to input it into our GCES for the extraction of added-value information mainly in terms of improving the quality of galaxies cataloguing.
An introductory text describing the data used is provided, as well as a list detailing the assumptions made when using this data. The summary of all results obtained is presented here along with an assessment in terms of the validity of the solution presented.
The following general assumptions and considerations were made when using the synthetic data:

- The realistic representativeness of the synthetic data set has only been pursuit up to the level of simulating some features of the instrument such as the gain ($\frac{electrons-per-pixel}{count-per-pixel}$) parameter.
- In general, IRAF (Image Analysis and Reduction Facility) package recommendations have been followed for the creation of synthetic images (e.g. source distributions).
- The Kron ellipses obtained through the artificial data sets have not included any adjustment parameter (e.g. K), contrary to what is normally done with real catalogues, where a number of manual steps conducted by experts is required in order to produce meaningful Kron ellipses to compute surface brightness. With this we intended to compare the performance of Kron ellipses versus our active contour on the same ground of no knowledge a priori or manual tuning regarding the computation of the contour.

4.2 Synthetic Data Set for the Photometric Cross Matching

The synthetic data is composed of tuples of fluxes with simulation of measurement in different bands, from ultraviolet to infra-red, and the photometric uncertainties associated with these measurements. This type of data is normally used in the so-called photometric catalogue, extensively used in the photometric redshift estimation problem. In order to obtain representative values of the measurements in each band, the COSMOS SED Grid is used as the basis for the creation of the synthetic multi-band photometric tuples.
The selection of an appropriate model for each data set is key for the correct development of a framework such as the one presented in this paper; therefore, careful consideration has been given to the selection of an appropriate SED templates library from which SED models are derived. There are two types of synthetic multi-band tuples which will simulate the two possibilities considered in the photometric cross-matching problem: match and non match. For the simulation of potential matches, the multi-band tuples were generated from a specific galaxy model of the SED template library by adding a synthetic random (from a Gaussian distribution) error to simulate the deviation from a perfect match between the data and the model, as in a real measurement. With this approach we can obtain a synthetic set of candidates of photometric multi-band matches, which will be analysed by our GCES. For the simulation of potential no matches, we randomly shuffle the synthetic ones.
The SWIRE SED and COSMOS SED template libraries from [Polletta et al., 2007] are the basis of our SED model library and they are also the source for the creation of synthetic tuples. The SWIRE SED library of templates contains 25 templates, including shapes such as elliptical, spiral, starburst, QSO, etc. The bands and associated transmission filters considered are: Subaru U, Subaru B, Subaru R, Subaru I, IRAC36, IRAC45, IRAC58 and IRAC 80, with an effective wavelength range from 3574.45$\mathring{A}$ to 78720$\mathring{A}$. The COSMOS SED library of templates, extracted from Le Phare package includes 31 SEDs which were used in [Ilbert et al., 2009]. It contains mainly SED templates for elliptical and spiral galaxies. The filters used are those corresponding to the bands considered in this research - CFHT bands: i, Ks and u; SUBARU bands: B, g, i, r, V, z; UKIRT band: J - and for an effective wavelength range from 3798$\mathring{A}$ to 21460$\mathring{A}$
The programme **sedtolib** included in Le Phare was used to build the different galaxy model libraries from the list of COSMOS SED files. In practical terms we used the following Le Phare programs: **sedtolib** - which builds the different galaxy libraries from a list of SED files with the goal of unifying the various SED original formats into a unique binary file ready for the next steps; **filter**- which gathers a list of filter response curves and applies some transformations according to the nature of the filters and **mag_gal**- which measures the magnitudes for the galaxy sample by computing these magnitudes for a set of given filters and an input SED library, and this can be computed at different redshifts.
The convolution of SED templates along the filter response curve and within a redshift range is described as follows (literal extract from Dr. Budavari notes of 1999):

$$F = \frac{\int S_\nu(\nu) r(\nu) \frac{1}{\nu} d\nu}{\int r(\nu) \frac{1}{\nu} d\nu}. \tag{4.1}$$

Where $S(\nu)$ is the spectrum and $r(\nu)$ is the filter's response function including the CCD's QE (Quantum Efficiency).
And using the transformation law $S_\nu = S_\lambda \cdot \frac{\lambda^2}{c}$ we obtain an algebraic equivalent of the previous equation:

$$F = \frac{\int S_\lambda(\lambda) r(\lambda) \lambda d\lambda}{c \int r(\lambda) \frac{1}{\lambda} d\lambda}. \tag{4.2}$$

Let f_λ be the rest frame spectrum of a galaxy at redshift z. Then, the spectrum $S(\lambda) = f_\lambda(\frac{\lambda}{1+z})$

and the effective flux in the band can be expressed as follows:

$$F = \frac{\int f_\lambda(\frac{\lambda}{1+z}) r(\lambda) \lambda d\lambda}{c \int r(\lambda) \frac{1}{\lambda} d\lambda}. \tag{4.3}$$

Re-binning to $log\lambda$ scale using the transformation $x \equiv log_q(\lambda/\lambda_0)$ yields the following:

$$\lambda = \lambda_0 q^x, \tag{4.4}$$

where $q = 1 + \Delta z$. Thus,

$$d\lambda = \lambda_0 q^x \ln(q) dx \tag{4.5}$$

and

$$\frac{\lambda}{1+z} = \lambda_0 q^{x-\xi} \tag{4.6}$$

where $1 + z = q^\xi$ and Δz defines the resolution of the logarithmic scale. The effective flux can now be read as:

$$F = \frac{\int f_\lambda(\lambda_0 q^{x-\xi}) r(\lambda_0 q^x) \lambda_0 q^x \lambda_0 q^x lnq dx}{c \int r(\lambda_0 q^x) \frac{1}{\lambda_0 q^x} \lambda_0 q^x \ln(q) dx}. \tag{4.7}$$

Defining the new function $f_x \equiv f_\lambda \frac{d\lambda}{dx}$ gives $f_x(x-\xi) = f_\lambda(\lambda_0 q^{x-\xi}) \lambda_0 q^{x-\xi} \ln(q)$. So the substitution yields:

$$F = \frac{\int f_x(x-\xi) q^\xi \lambda_0 q^x r(x) dx}{clnq \int r(x) dx}, \tag{4.8}$$

which can be expressed as:

$$F = \frac{1+z}{c \ln(1+\Delta z)} \cdot \frac{\int f_x(x-\xi) \lambda(x) r(x) dx}{\int r(x) dx}. \tag{4.9}$$

We build up a grid of the SED COSMOS models convolved with the filters enumerated above and this convolution was done for a range of redshifts from 0.04 to 6. In practical terms we performed the following steps:

- Run Le Phare **sedtolib** for the COSMOS SED templates - Ell1_A_0, Ell2_A_0, Ell3_A_0, Ell4_A_0, Ell5_A_0, Ell6_A_0, Ell7_A_0, S0_A_0, Sa_A_0, Sa_A_1, Sb_A_0, Sb_A_1, Sc_A_0, Sc_A_1, Sc_A_2, Sd_A_0, Sd_A_1, Sd_A_2, Sdm_A_0, SB0_A_0, SB1_A_0, SB2_A_0, SB3_A_0, SB3_A_0, SB4_A_0, SB5_A_0, SB6_A_0, SB7_A_0, SB8_A_0, SB9_A_0, SB10_A_0, SB11_A_0.
- Run Le Phare **filter** for the filters - CFHT u*, Subaru Bj, Subaru Vj, Subaru g+, Subaru r+, Subaru i+, Subaru z+, CFHT i*, CFHT Ks and UKIRT J.
- Run the c++ programme **convipoltest** from **Photoz** obtaining a matrix of convolved fluxes for the above SEDs, filters and in a defined range of redshifts from 0 to 6.
- Run Le Phare **mag_gal** to derive the matrix of convolved magnitudes for the whole SED COSMOS model library
- Run a dedicated script in Matlab to convene the format of the SED COSMOS model library into a grid of parameters T, z, m as described in Chapter 3.

A Matlab function was coded to create the synthetic photometric tuples. In this function, the input is composed of a grid of SED models, specific values for T and z which select a specific type of galaxy in the model, a selected value for the mean and deviation of the simulated uncertainties (sigma) and

a smooth factor. The output is composed of a tuple of AB magnitudes and corresponding sigma values. With this function, we simulate realistic and representative measurement errors which in reality explain the lack of perfect match between the model and the data. For the computation of both output values, we use the Gauss Matlab function which, after a thorough analysis and validation (described in Chapter 3) has proven to be a valid function for this objective.
For the validation purposes of this chapter, representative values for measurement uncertainties- $\sigma_\mu = 0.3$ and $\sigma_\sigma = 0.05$ were chosen for the case of true match and $\sigma_\mu = 0.8$ and $\sigma_\sigma = 0.5$ for the case of wrong match. σ_μ and σ_σ are the mean and the deviation of the simulated measurements yielded from a chosen element of the SED grid. The following set of synthetic true match 10-bands tuples were created:

- A set of 1000 synthetic 10-band tuples created from an elliptical galaxy (Ell7_A_0) and a redshift z =1.10.
- A set of 1000 synthetic 10-band tuples created from a spiral galaxy (SB1_A_0) and a redshift z = 3.90.

Note that the wording "good match", when found along the text of this thesis refers to a true match; conversely the naming "bad match" refers to a wrong match.
As indicated above, the synthetic tuples for cases of bad matches were created in a manner similar to the tuples for good matches, but having been obtained, the components of the whole matrix of tuples were shuffled in the following manner: first we swap for each synthetic tuple created the components 1 and 2 by the components 9 and 10 respectively. Similarly we swap the component 3 and 4 of each tuple by the components 5 and 6. Once this is done, let N be the number of synthetic tuples of 10 bands created. Then we swap the components 1 to 5 of the N/2 tuples with the components 6 to 10 of this very half set of tuples. In this manner we can evaluate the influence in the matching results for this shuffle approach when creating synthetic no match tuples. The following synthetic bad match 10-bands tuple was created:

- A set of 1000 synthetic 10-band tuples created from a spiral galaxy (SB8__0) and redshift z = 4.
- A set of 1000 synthetic and shuffled 10-band tuples created from a spiral galaxy (SB2_A_0) and a redshift z = 2.05.

The following sections describe the results obtained from the implementation of the photometric cross-matching approach with these sets of synthetic photometric multi-band tuples. The number of synthetic data set created ensured a statistical representativeness of the results.

4.3 Synthetic Data Set for the Source Contour and Labelling

The IRAF package was used to create artificial images to serve as the basis for the validation of the surface brightness computation. The description of the IRAF functionalities and main parameters used were extracted from the interactive help offered by the Ureka platform where IRAF was installed and used. Ureka is a collection of useful astronomical software that is mostly written in Python and centred around IRAF. It has been deprecated as of 4/26/2016. However at that time all the usage needed for the purpose of this research was done. Therefore there was no need to use the new **AstroConda** solution.
The computational cost for an accurate simulation of effects such as the sampling of the pixel, the effect of atmospheric seeing, object appearance, luminosity functions and noise should not be underestimated. IRAF includes many approximations and algorithms to make this possible to as high a degree of accuracy as practical.

New images are created with specific dimensions. A constant parameter gives the images background value. Noise consists of Gaussian and Poisson components. Synthetic objects in the image are specified by a position, magnitude, model, scale, axial ratio and position angle. The object lists specify the intrinsic shapes.
IRAF offers a high degree of flexibility in defining the object models. The models are set either as one-dimensional radial intensity or with a cumulative flux profile or by an image template. It is noted in the Ureka-IRAF help that intensity profiles are to be preferred to avoid artefacts in the conversion from cumulative flux.
The Sérsic (*sersic*), exponential disk (*expdisk*) and De Vaucouleurs (*devauc*) profiles are common models for spirals and elliptical galaxies.
We integrate the models of the objects over the size of the image in pixels. This is done by sub-sampling, dividing up a pixel into smaller pieces named sub-pixels. The galaxies are sub-sampled less since they will have less rapidly changing profiles and are convolved by the PSF (Point Spread Function). While the stars are computed a few times, we need to calculate each galaxy separately. Thus we can achieve greater precision in the stars than in the galaxies.
This model integration step can be very expensive in computation time. IRAF proposes a more flexible approach by adjusting the sub-sampling parameters depending on the need. For example, the number of sub-pixels in each dimension in each pixel could decrease linearly with the distance from the profile centre. With this, we can reach a fair trade-off for the degree of accuracy desired against a lower execution time.
A significant difference to remark here is that contrary to the stars, galaxies have different profiles, ellipticity and position angles. Another difference is that we need to convolve the galaxy models with the PSF; i.e. the shapes are defined before seeing. The PSF convolution must also be sub-sampled, and the number of operations required for each convolution operation is comparable to the number of pixels in the PSF for each galaxy sub-pixel. Thus, computing seeing convolved, well sub-sampled, large galaxy images is the most demanding task of all.
The PSF used when convolving galaxies is usually truncated. This truncation implies indeed the loss of a small amount of the flux in the wings, but it alleviates the computational cost considerably.
The following text with grey background is literally taken from the IRAF manual and the Ureka IRAF installed help.

Two commands were used for the creation of the synthetic images and extraction of the magnitudes: **gallist** and **mkobjects**. The following list indicates the parameters used between both commands:

- **gallist**: name of the output text file for the x and y coordinates, magnitudes, profile types, half-flux radii, axial ratios and position angles of the artificial galaxies.
- **ngals**: number of galaxies in the output galaxies list.
- **spatial**: type of spatial distribution for the galaxies. The following two types were considered in this research:
  - *uniform*: the galaxies are uniformly distributed between xmin, xmax, ymin, ymax ($1 to 512$ for each axis).
  - *Hubble*: the galaxies are distributed around the center of symmetry xcenter and ycenter according to a Hubble density law of core radius (50 pixels) and background density base (0.0).
- **luminosity**: type of luminosity distribution for the galaxies. The type considered in this research is *schechter* - the galaxies are distributed according to a schechter luminosity function with characteristic magnitude *mstar* and power law exponent *alpha* between *minmag* (-7) and *maxmag* (0). The following parameters are considered for the schechter luminosity function:

  - *mzero*: Magnitude zero point for schechter luminosity function. The value used is 15.
  - *alpha*: The power law exponent of the schechter luminosity function. The default value is that determined by schechter for nearby galaxies.
  - *mstar*: The characteristic magnitude of the schechter luminosity function. The value used is $-21,41$.

egalmix: the fraction of the galaxies that are "ellipticals" represented by a de Vaucouleurs surface brightness law as opposed to "spirals" represented by an exponential disk surface brightness law.

- **z**: Minimum redshift for power law distributed galaxies. This is the redshift assigned galaxies of magnitude *minmag*. The redshifts are assumed proportional to the square root of the apparent luminosity; i.e. the luminosity distance proportional to redshift. The redshift is used to compute the mean apparent size of the galaxies according to $(1+z)^2/z$. The value used here for minimum redshift is 0.05.
- **title**: image title to be given to the images. Maximum of 79 characters.
- **ncols and nlines**: Number of columns and lines. The values used here are 512 for both columns and lines.
- **background**: Default background and poisson noise background. This is represented in data numbers with the conversion of photons determined by the gain parameter. The value used is 1000.
- **objects**: list of objects files. The number of object files must match the number of input images. The object files contain lines of object coordinates, magnitudes and shape parameters. The values used are 5,10,50,100,1000 for the various cases of creation of synthetic images and galaxies inside them.
- **distance**: Relative distance to be applied to the object list coordinates, magnitudes and scale sizes. This factor is divided into the object coordinates, after adding the offset factors, to allow expanding or contracting about any origin. The magnitudes scale as the square of the distance and the sizes of the galaxies scale linearly. This parameter allow changing image sizes and fluxes at a given seeing and sampling with one value. The value used is 1.
- **exptime**: Relative exposure time. The object magnitudes and background levels are scaled by this parameter. This is comparable to changing the magnitude zero point except that it includes changing the background. The value used is 1
- **magzero**: Magnitude zero point defining the conversion from magnitudes in the object list to instrumental/image fluxes.
- **gain**: Gain in electrons per data number. The gain is used for scaling the read-noise parameter, the background and in computing Poisson noise. The value used are 1, 5 and 10.
  From general CCDs (Charged-Coupled Devices) bibliography, it is clear that there are a few steps involving hardware and software when reading the value of a pixel, based on how electrons move via electrodes placed on the silicon surface. These steps can be summarized as follows:

1 Electrons transferred to "amplifier" (really a capacitor). Units are columbs.
2 The voltage induced by this charge is measured. Units are volts
3 An Analogue-To-Digital(A/D) unit converts the voltage into some other voltage which may have discrete values. Units still volts.
4 The voltage is converted into a number which is passed from the hardware to the computer software as the pixel's value. Units are counts, also named "Data Numbers" (DN) or Analogue-to-Digital Units (ADUs).

In both steps 3 and 4, one can scale the result by any arbitrary factor and the relative pixel values will remain the same. Some software allows the user to modify the scaling factor dynamically; others have a fixed setting.
The end result is that there is some "fudge factor" which relates the initial number of electrons in a pixel to the final number of counts reported by camera software. The ratio of these two numbers is the gain of the camera: $gain = \frac{number-of-electrons-per-pixel}{number-of-counts-per-pixel}$

- **rdnoise**: Gaussian read-noise in electrons. For new images this applies to the entire image while for existing images this is added only to the objects. The value used are 0 and 10 depending on the case of creation of synthetic data without or with read-noise. This value is the one recommended in IRAF.
- **poisson(yes/no)**: Add poisson photon noise?. For new images this applies to the entire image, while for existing images this is only applied to the objects. Note that in the latter case the background parameter is added before computing the new value and then subtracted again. The values used are *yes* and *no* depending on the case of synthetic data with or without poisson noise.
- **magnitude**: the object magnitude is converted to instrumental fluxes as follows:

$$flux = \frac{exptime}{distance^2} \cdot 10^{(-0.4 \cdot (magnitude - magzero))} \tag{4.10}$$

In practical terms the following set of synthetic images and galaxies were considered:

- Effect of crowding areas: Various number of galaxies - 50, 100 and 1000 - for the size of the image-512, in Uniform spatial distribution and without noise, using the same gain value for the three cases (default gain value of 10 as indicated in the Ureka package). Using the same value for the gain will ensure no impact in the comparative assessment of the various crowding factors. This set of synthetic data will allow an evaluation of the behaviour of our active contour algorithm for the cases of crowded and non crowded images under ideal- no noise- conditions.
- Effect of Poisson noise with two distributions and two crowding factors: we execute active contour algorithm for two cases - 100 Hubble galaxies with very low gain value - (1) - and Poisson noise and 425 Uniform galaxies with very high gain value - (95) - and Poisson noise. This set of synthetic data will allow an evaluation of the robustness of our active contour algorithm against various conditions of Poisson noise, gain and galaxy distributions.
- Effect of both Poisson and read-noise, various gain and galaxy distributions and crowding areas: we execute our surface brightness algorithm for the following six cases - 100 Hubble galaxies with gain value 1, read-noise value 10 and Poisson noise; 20 Uniform galaxies with gain value 3, read-noise value 20 and Poisson noise; 50 Uniform galaxies with gain value 10, read-noise value 20 and Poisson noise; two sets of 425 Uniform galaxies with gain values 10 and 30 respectively, read-noise value 10 and Poisson noise and finally 425 Uniform galaxies

with gain value 70 no read-noise and Poisson noise. This synthetic data intends to simulate the expected real data context covering a realistic range for gain value and read-noise. The number of 425 galaxies is proportional to the number of real galaxies for the dimension of the image. The Uniform distribution is the type of distribution recommended by IRAF for validation purposes. An assessment on the capabilities of the active contour algorithm under severe conditions (worst case scenario) was performed as part of the validation evaluation.

The use of SExtractor package to extract the apparent magnitudes from synthetic images was done with a reduced set of synthetic data and using the default configuration file of SExtractor. A gain value of 3 was used.
IRAF magnitudes are the true values created from a predefined luminosity profile and the image is composed from that profile. However, SExtractor computes the magnitudes from an input image and, for the same input, the output (magnitudes) can vary considerably depending on the configuration file, which typically requires the knowledge of an expert. Thus, for the validation of our GCES contouring algorithm we use the IRAF magnitudes as the true reference.

4.4 Photometric Cross Matching Validation

Figures - 4.1 to 4.17 present the results of the photometric cross matching validation with the synthetic data described in Section 4.2. These results show the following perspectives of the photometric cross matching problem for the two simulated cases of true matches and wrong matches (from shuffled SEDs):

- The distribution of decimal logarithms of photometric Bayes Factor ($log_{10}PhBF$) for the synthetic tuples.
- The distribution of redshifts for the best SED fitting of the synthetic tuples when calculating the photometric Bayes Factor.
- The Best SED fitting of the synthetic tuple with the highest value of the photometric Bayes Factor.
- The evolution of the values of $log_{10}PhBF$ with increasing added noise.

As expected, the distribution of $log_{10}PhBF$ values is shifted towards the left, i.e. negative values, for the cases of wrong matches and towards the right, i.e positive values larger than unity for the simulated true matches. Regarding the redshift and SED template distributions there is a good correspondence, for the cases of matches, of the results which concentrate around the predefined values in the synthetic data set. Conversely, the cases of no matches show the expected dispersion in these parameters compared to the predefined in the synthetic data set.
Figure 4.1 shows the histogram of the decimal logarithm of the Bayes Factor when implementing the photometric cross matching solution with the 1000 synthetic tuples created from the spiral galaxy $SB1\_A\_0$ of the SED grid described in section 4.2. This dataset was created to validate the case of true matches. We observe that the vast majority of result correspond to clear matches (around 90%) while around 8% of the results yield negative Bayes Factor, that is, those synthetic tuples are no matches.
Figure 4.2 represents the best SED fitting for the tuple that yielded the maximum value of Bayes Factor $PhBF$ in the photometric cross matching problem. We observe that the fitting is almost perfect, that is, the value of magnitude of that synthetic tuple in each band is very close to the corresponding values of the SED that the system identified from the SED Grid as the best fit for that tuple in question. We observe slight variations in the parameters of redshift z and SED template T achieved for the best SED of the maximum PhBF, which are 3.68 and 22 in front of 3.90 and

21 of the SED selected from the Grid to create the 1000 synthetic tuples. This is due to the errors introduced to create realistic synthetic dataset.
Figure 4.3 shows the distribution of redshift values z yielded for the best SED fit achieved with each of the 1000 tuples as part of the potometric cross matching problem. We observe that the redshift values are concentrated around the value of redhift of 3.9 which correponds to the redshift of the SED from which we created the 1000 synthetic tuples.
Figure 4.4 shows the distribution of SED template types T obtained for the best SED fit achieved for each of the 1000 synthetic tuple. This SED fitting step is done as part of the photometric cross matching problem. We observe that the T values concentrate around 21 which correspond to the spiral galaxy $SB1\_A\_0$ of the SED Grid. With this, we observe that there is a very good correspondence of the results on a representative number of synthetic data set with the expectation which is firstly to identify the majority of the cases as matches and secondly to find the best SED fit around the parameters of the SED from which the synthetic data set was created.

Figure 4.5 shows the histogram of the decimal logarithm of the Bayes Factor when implementing

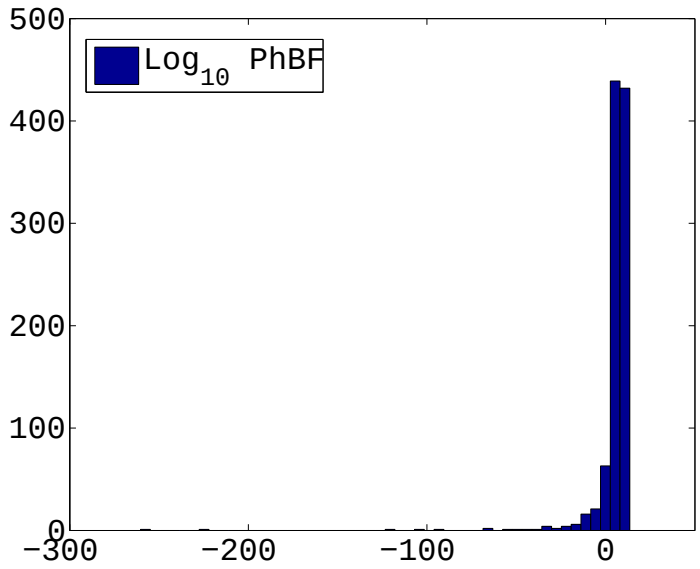

Figure 4.1: Distribution of photometric Bayes Factor logarithmic values for 1000 simulated matches tuples of SB1_A_0 galaxy with redshift 3.90. It is observed, as expected, that the majority of values concentrate around high Bayes Factor because these are simulated matches.

the photometric cross matching solution with the 1000 synthetic tuples created from the elliptical galaxy $Ell7\_A\_0$ of the SED grid described in section 4.2. This dataset was created to validate the case of true matches. We observe that the vast majority of result correspond to clear matches (around 85.6%) while around 13.5% of the results yield negative Bayes Factor, that is, those synthetic tuples are no matches.
Figure 4.6 represents the best SED fitting for the tuple that yielded the maximum value of Bayes Factor $PhBF$ in the photometric cross matching problem. We observe that the fitting is almost perfect, that is, the value of magnitude of that synthetic tuple in each band is almost coincident with the corresponding values of one of the SEDs that the system identified from the SED Grid as

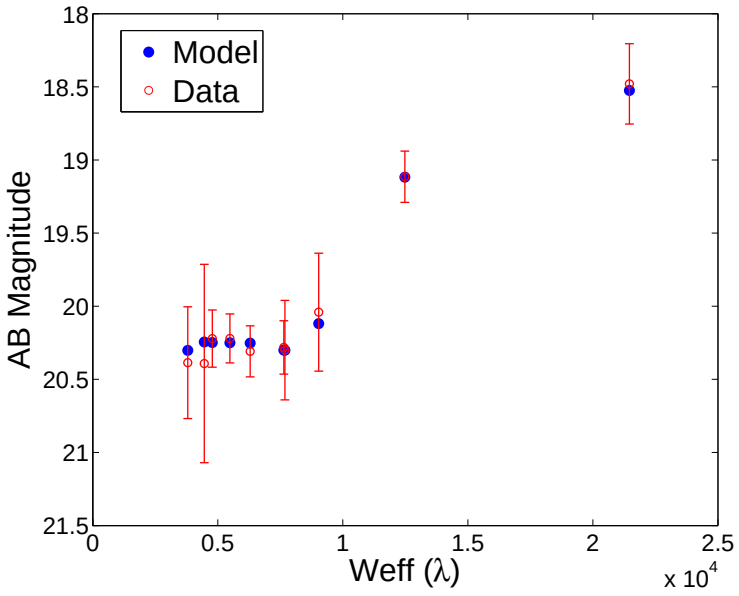

Figure 4.2: Best SED fitting for the maximum value of Photometric Bayes Factor ($3.9059 * 10^{13}$). It corresponds to a spiral SB2_A_0 galaxy with redshift 3.68

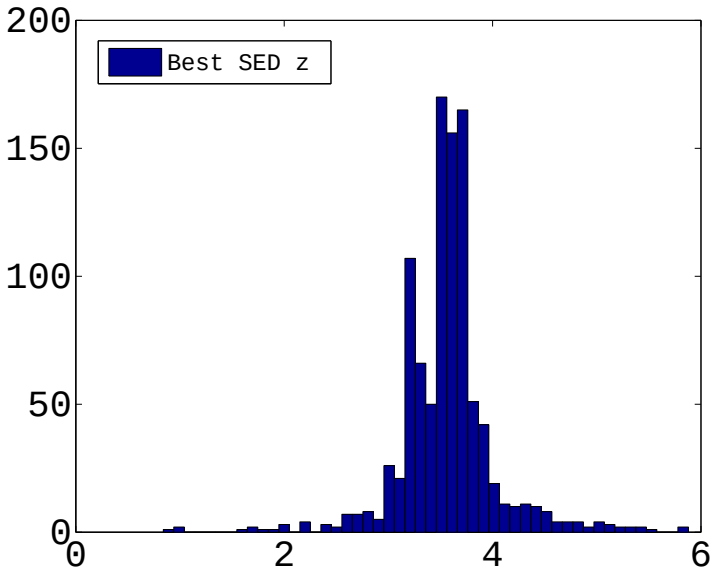

Figure 4.3: Distribution of redshift for the best SED fitting with the synthetic matched tuples of a SB1_A_0 galaxy with redshift 3.90.

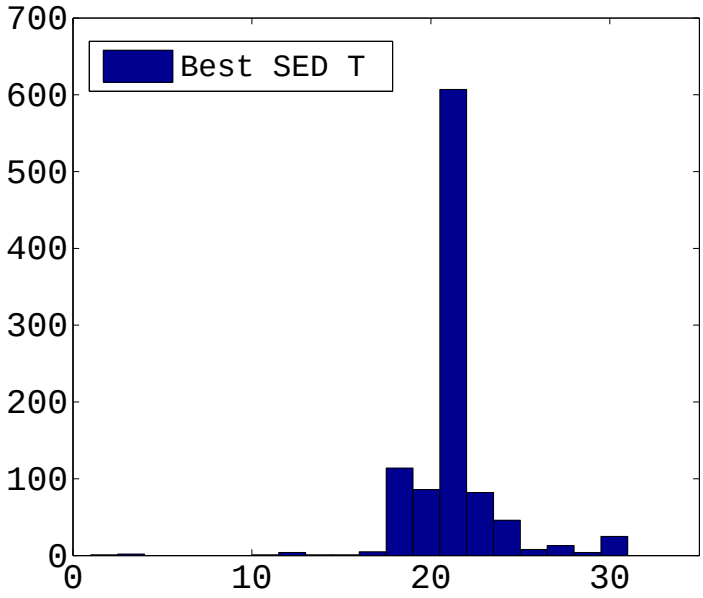

Figure 4.4: Distribution of SED templates for the best SED fitting with the synthetic matched tuples of a SB1_A_0 galaxy with redshift 3.90.

the best fit for that tuple in question.We observe slight variations in the parameters of redshift z and SED template T achieved for the best SED of the maximum PhBF, which are 0.96 and 6 in front of 1.10 and 7 of the SED selected from the Grid to create the 1000 synthetic tuples. This is due to the errors introduced to create realistic synthetic dataset.
Figure 4.7 shows the distribution of redshift values z yielded for the best SED fit achieved with each of the 1000 tuples as part of the potometric cross matching problem. We observe that the redshift values are concentrated around the value of redhift of 1.1 which correponds to the redshift of the SED from which we created the 1000 synthetic tuples.
Figure 4.8 shows the distribution of SED template types T obtained for the best SED fit achieved for each of the 1000 synthetic tuple. This SED fitting step is done as part of the photometric cross matching problem. We observe that the T values concentrate around 7 which correspond to the spiral galaxy $Ell7\_A\_0$ of the SED Grid. With this, we observe that there is a very good correspondence of the results on a representative number of synthetic data set with the expectation which is firstly to identify the majority of the cases as matches and secondly to find the best SED fit around the parameters of the SED from which the synthetic data set was created.

Figure 4.9 shows the histogram of the decimal logarithm of the Bayes Factor when implementing the photometric cross matching solution with the 1000 synthetic tuples created from the spiral galaxy $SB2\_A\_0$ of the SED grid described in section 4.2. This dataset was create to validate the case of wrong matches. We observe that the vast majority of result correspond to clear matches (around 7.6%) while around 92.4% of the results yield negative Bayes Factor, that is, those synthetic tuples are no matches.
Figure 4.10 represents the best SED fitting for the tuple that yielded the minimum value of Bayes Factor $PhBF$ in the photometric cross matching problem. We observe that the fitting absolutely

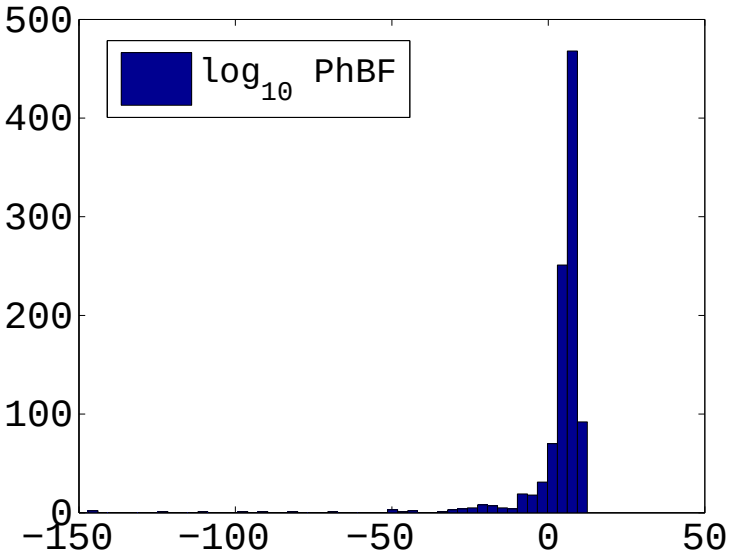

Figure 4.5: Distribution of photometric Bayes Factor logarithmic values for 1000 simulated mathes tuples of Ell7_A_0 galaxy with redshift 1.10. It is observed, as expected, the majority of values around high Bayes Factor because these are simulated matches.

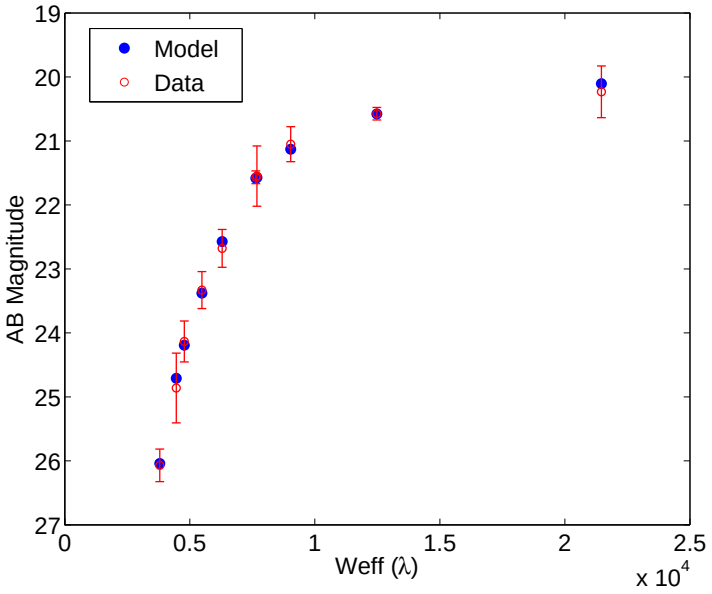

Figure 4.6: Best SED fitting for the maximum value of Photometric Bayes Factor ($3.1328 * 10^{12}$). It corresponds to an elliptic Ell6_A_0 galaxy with redshift value of 0.96.

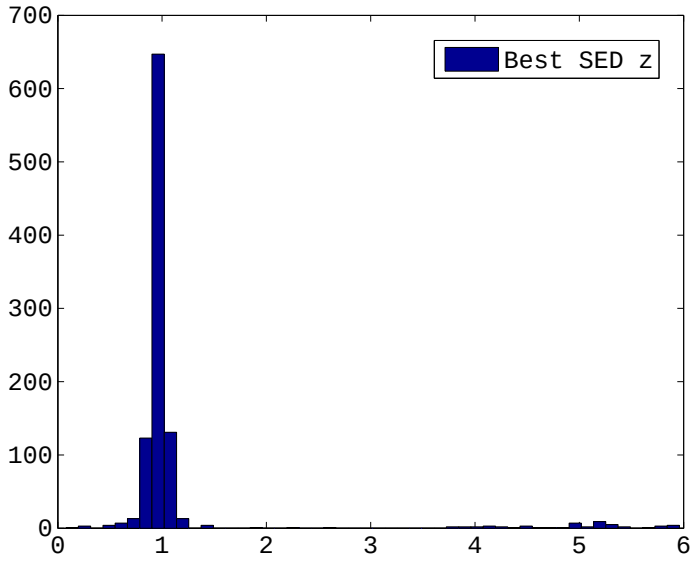

Figure 4.7: Distribution of redshift for the best SED fitting with the synthetic matched tuples of a ELL7_A_0 galaxy with redshift 1.10.

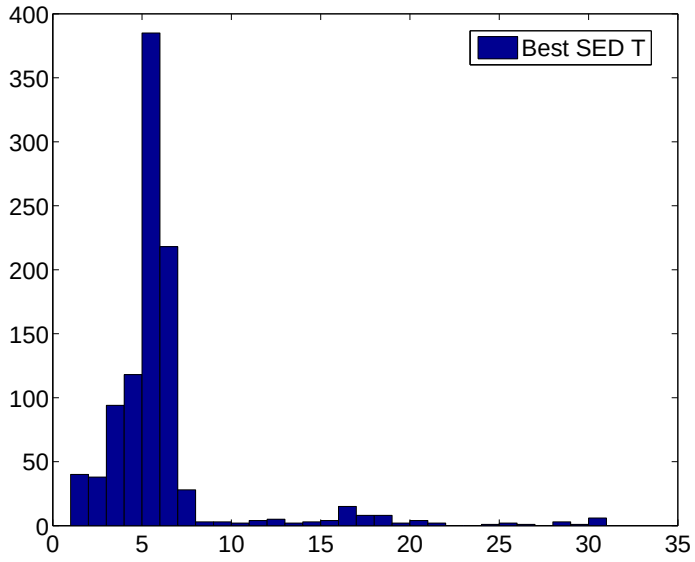

Figure 4.8: Distribution of SED templates for the best SED fitting with the synthetic matched tuples of a Ell7_A_0 galaxy with redshift 1.10.

poor, that is, the value of magnitude of that synthetic tuple in each band shows large difference, beyond the errorbars, with the corresponding values of the SEDs that the system identified from the SED Grid as the best possible fit for that tuple in question. The parameters z and T for this best possible SED fit are 2.52 and 1 which coincidently is close in the redshift 2.5 but largely different in the SED template type, 22 for the SED selected from the Grid to create the 1000 synthetic tuples. This is due to the errors introduced to create realistic synthetic dataset. In addition to that and for the case of synthetic no matches, the synthetic tuples of the data set were heavily shuffled to ensure proper representativeness of the case of no match.
Figure 4.11 shows the distribution of redshift values z yielded for the best SED fit achieved with each of the 1000 tuples as part of the potometric cross matching problem. We observe that the redshift values are heavily distributed, such that the histogram is approaching a uniform distribution more than a normal one as in the cases of true matches.
Figure 4.12 shows the distribution of SED template types T obtained for the best SED fit achieved for each of the 1000 synthetic tuple. This SED fitting step is done as part of the photometric cross matching problem. We observe that the T values concentrate around a value of T largely different, 30 from the one of the SED which created the synthetic dataset, 22. Therefore and, as expected for the case of wrong matches, we observe that there is no good correspondence between the synthetic data set and the SED Grid against which we explore the cross-matching problem.

Figure 4.13 shows the histogram of the decimal logarithm of the Bayes Factor when implementing

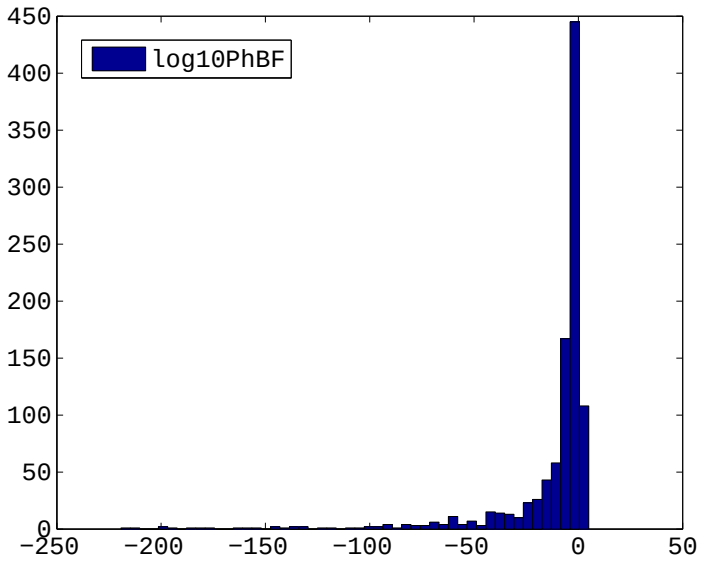

Figure 4.9: Distribution of photometric Bayes Factor logarithmic values for 1000 simulated wrong matches tuples (shuffled SEDs) of SB2_A_0 galaxy with redshift 2.05. It is observed, as expected, the majority of values around low Bayes Factor because these are shuffled SEDs.

the photometric cross matching solution with the 1000 synthetic tuples created from the spiral galaxy $SB8\_A\_0$ of the SED grid described in section 4.2. This dataset was create to validate the case of wrong matches. We observe that the vast majority of result correspond to clear matches (around 5.2%) while around 94.80% of the results yield negative Bayes Factor, that is, those synthetic tuples

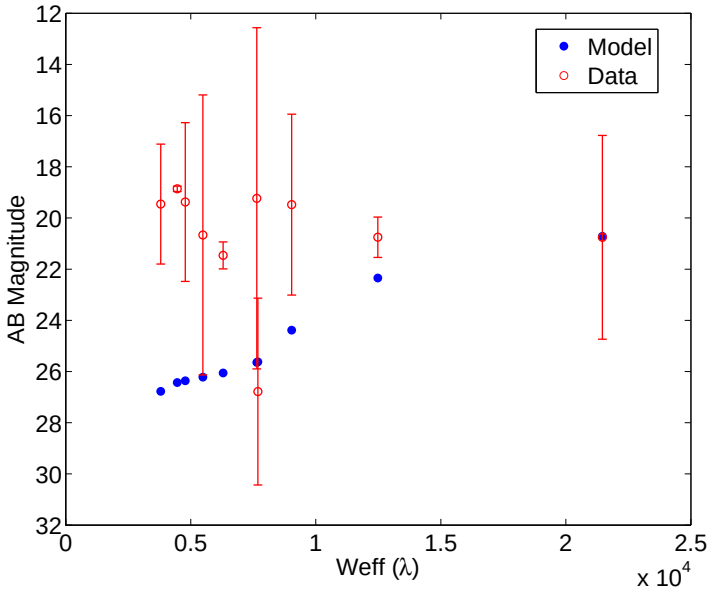

Figure 4.10: SED fitting for the minimum value of Photometric Bayes Factor ($5.5361 * 10^{-220}$). It corresponds to a spiral S0_A_0 galaxy with redshift value of 2.52.

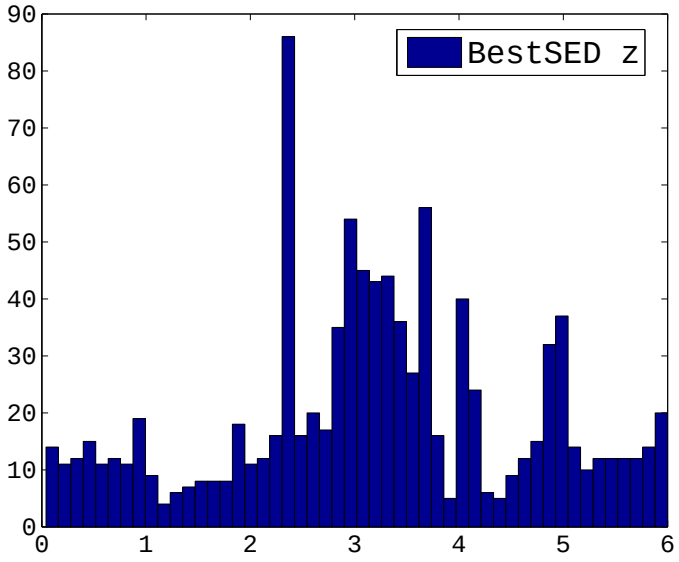

Figure 4.11: Distribution of redshifts for the best SED fitting with the synthetic shuffled SED tuples of a SB2_A_0 galaxy with redshift 2.05.

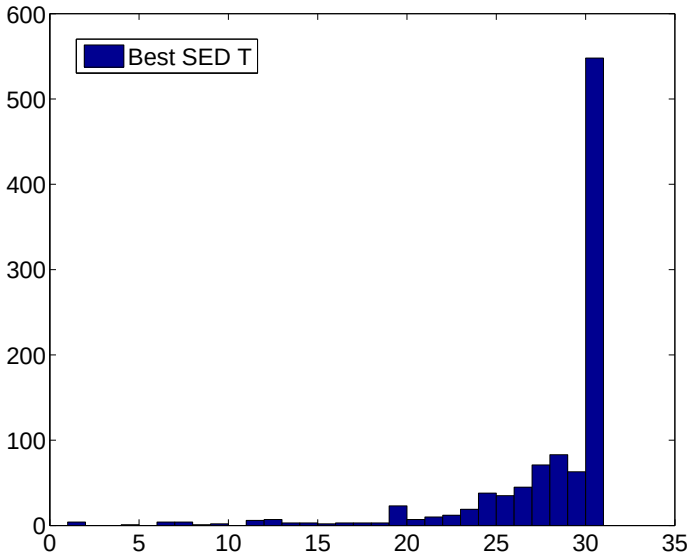

Figure 4.12: Distribution of SED templates for the best SED fitting with the synthetic no matched tuples of a SB2_A_0 galaxy with redshift 2.05.

are no matches. It is important to note here that the cases
Figure 4.14 represents the best SED fitting for the tuple that yielded the minimum value of Bayes Factor $PhBF$ in the photometric cross matching problem. We observe that the fitting absolutely poor, that is, the value of magnitude of that synthetic tuple in each band shows important differences, beyond the errorbars for three bands, with the corresponding values of the SEDs that the system identified from the SED Grid as the best possible fit for that tuple in question. The parameters z and T for this best possible SED fit are 4.12 and 31 which coincidently is close in the redshift 4 but clearly different in the SED template type, 28 for the SED selected from the Grid to create the 1000 synthetic tuples. This is due to the errors introduced to create realistic synthetic dataset. In addition to that and for the case of synthetic no matches, the synthetic tuples of the data set were heavily shuffled to ensure proper representativeness of the case of no match.
Figure 4.15 shows the distribution of redshift values z yielded for the best SED fit achieved with each of the 1000 tuples as part of the potometric cross matching problem. In this case we observe that the redshift values are in concentrated in three peaks of z values of around 2, 3.7 and 4, thus we observe a trend towards a multi variant distribution which can be explained as a consequence of the shuffling of the synthetic tuples leading to a dispersion in the SED parameters.
Figure 4.16 shows the distribution of SED template types T obtained for the best SED fit achieved for each of the 1000 synthetic tuple. This SED fitting step is done as part of the photometric cross matching problem. We observe that the T values concentrate around a value of T largely different, 31 from the one of the SED which created the synthetic dataset, 8. Therefore and, as expected for the case of wrong matches, we observe that there is no good correspondence between the synthetic data set and the SED Grid against which we explore the cross-matching problem.

Figure 4.17 shows the result of computing the photometric Bayes Factor for simulated data created from an elliptical SWIRE SED template using different values of photometric uncertainty to simulate real measurements with different degrees of error; clearly, as expected, high values of uncertainty in

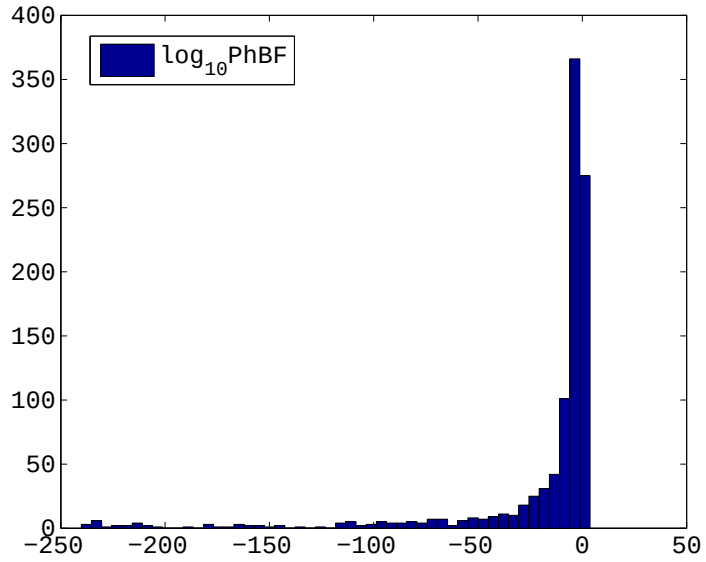

Figure 4.13: Distribution of photometric Bayes Factor logarithmic values for 1000 simulated no matched tuples of SB8_A_0 galaxy with redshift 4. It is observed, as expected, the majority of values around low Bayes Factor because these are simulated no matches.

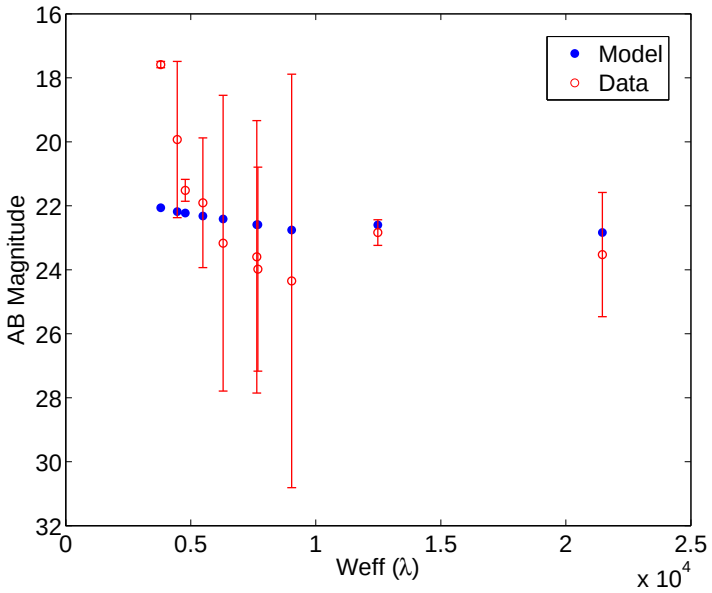

Figure 4.14: SED fitting for the minimum value of Photometric Bayes Factor ($6.9813 * 10^{-241}$). It corresponds to a spiral SB11_A_0 galaxy with redshift value of 4.12.

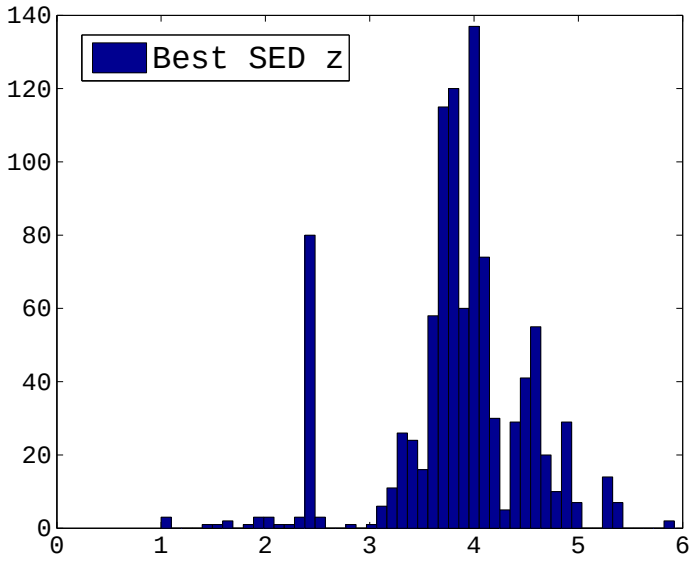

Figure 4.15: Distribution of redshift for the best the SED fitting with the synthetic shuffled SED tuples of a SB8_A_0 galaxy with redshift 4.

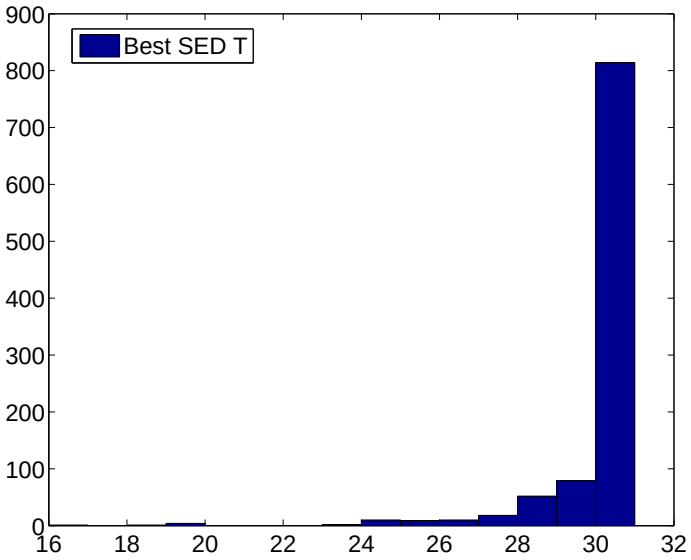

Figure 4.16: Distribution of SED templates for the best SED fitting with the synthetic no matched tuples of a SB8_A_0 galaxy with redshift 4.

the data produces poor results in terms of best fit and of matches, and little uncertainty in the data produces a good quality fit and high values of the photometric Bayes Factor for simulated matched samples.

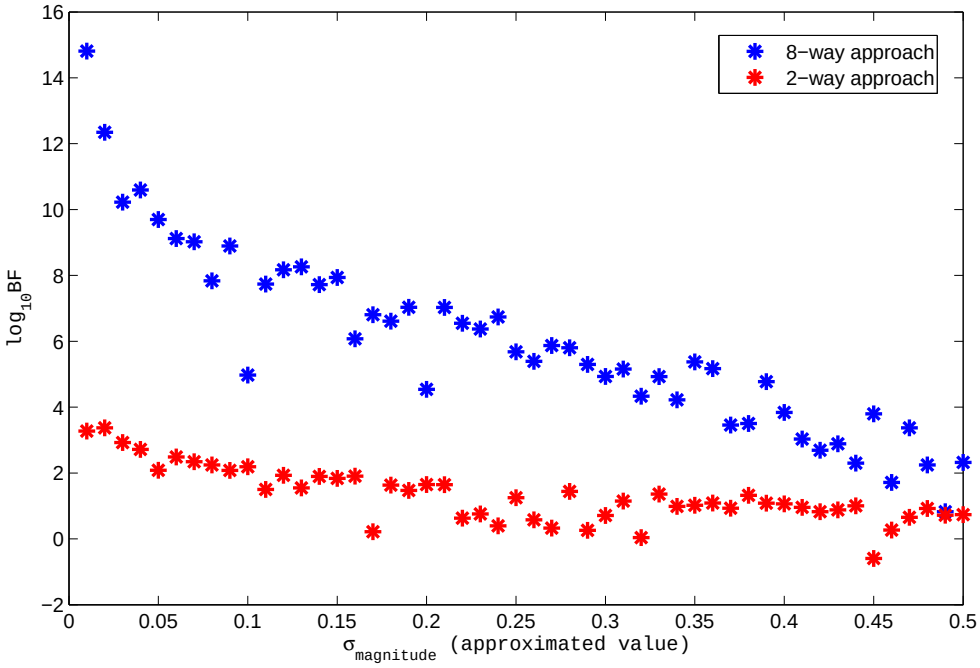

Figure 4.17: Solution of the Bayes Factor with increasing noise for simulated data created from two groups of four bands: one group encompasses the Subaru bands U, B, R, I and the other group encompasses the Spitzer bands IRAC36, IRAC45, IRAC58, and IRAC85, and eight-way matches (named n-way approach in the figure) created from Subaru bands U, B, R, I and Spitzer IRAC bands IRAC36, IRAC45, IRAC58 and IRAC85 without grouping them per instrument. The samples are created from a specific elliptic, Ell5, and SWIRE SED templates of redshift 2.68, adding measurement errors to simulate real measurements.

4.5 Source Contour Extraction Validation

In order to validate the approach proposed here, a set of artificial images and their derived catalogues have been created so that the conditions of the validation are properly controlled. We have used the **artdata** package of the IRAF tool to create various images with artificial galaxies. As indicated at the beginning in Section 4.3, our source contour algorithm was run on a set of artificial data sets under a controlled environment; i.e., features of these images such as signal-to-noise, spatial distribution of the galactic sources and background brightness were part of the design of these artificial data sets. Similarly, the catalogues obtained from these artificial images contained precise astrometric information in terms of coordinates (celestial and pixel), Kron ellipses parameters, etc, as well as photometric information in terms of brightness (flux, apparent magnitude), background,

zero point, etc.
The parameters λ_1, λ_2, initial radius and maximum number of iterations were explored and assessed in order to identify the best values for our contour solution. This exercise was conducted on an ideal image of 50 galaxies in a Uniform distribution with gain 10 and no noise, shown in Figure 4.18. The values of the parameters were chosen such that the range covered all possible results. Figures 4.18 and 4.19 present the result achieved with the best configuration for this type of synthetic image: $\lambda_1 = 16$, $\lambda_2 = 4$, $r = 3$ and $Itermax = 15$. With this result we confirm the adequacy of our source contour definition algorithm to its use in our GCES system. We also demonstrated that our current implementation needs further refinement for the cases of blended sources typically found in crowded areas of the sky. This refinement is proposed as a further line of research.

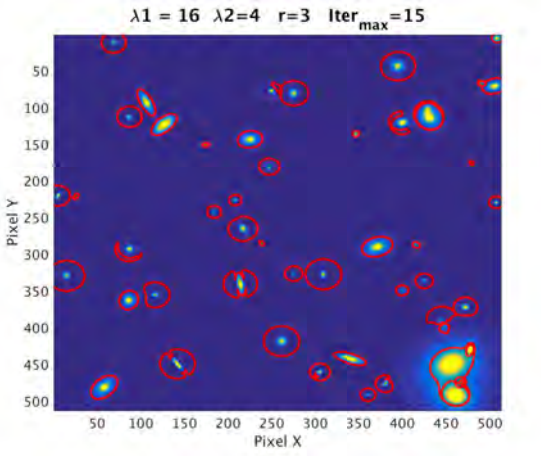

Figure 4.18: GCES Contour (in red)

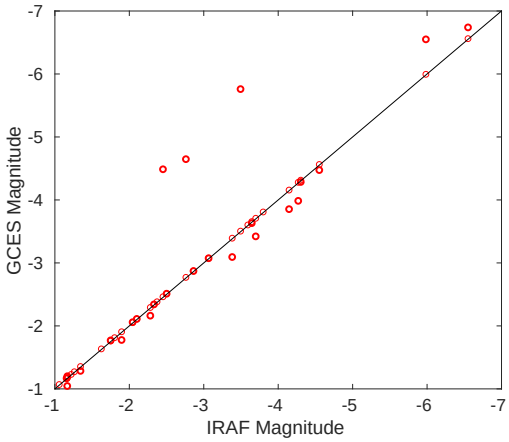

Figure 4.19: AB Magnitude values of GCES and IRAF catalogue

Figures 4.20, 4.21 and 4.22 illustrate the outcome of running our GCES contour algorithm on a synthetic image of 50 uniform sources with no noise, as well as two representative cases of good and bad behaviour of our GCES contouring.
Figure 4.20 represents the values of magnitude from our GCES system versus the real magnitudes yielded by IRAF. The black line in the diagonal of this figure is the reference for the ideal case of GCES magnitude being the same as the one from IRAF. The image used here is the one of figure 4.18, where we observe that except very few blended sources, the majority are isolated. Moreover, noise is not included in this image. Thus, the GCES contours are obtained for an ideal, non realistic scenario. Figure 4.20 shows a very good alignment of the GCES magnitudes with the IRAF values, which indicates that our GCES system is capable of automatically extract the contour of the source with a very good degree of accuracy. The few cases where this alignment is cleary compromised corresponds to blended sources. The current version of our GCES contour algorithm does not handle properly these cases. A future line of work is presented in Chapter 6 addressing this limitation. The GCES contour for two representative cases, indicated with blue asterix in Figure 4.20 are shown in Figures 4.21 and 4.22. Figure 4.21 presents one of the best cases of performance of our GCES contour algorithm. We observe in this case, that the GCES contour covers all the area of the source. The difference between both magnitude values: GCES and IRAF, is of order of 10^{-4}. Therefore, our GCES contouring solution is very reliable for cases of isolated sources with no apparent noise. Conversely, Figure 4.22 represents one case of poor alignment in magnitude values between the true reference provided by IRAF and our GCES system. The reason for this poor behaviour is found in the two blended sources encountered in that figure (pixel coordinates marked with black crosses),

for which our GCES system drew one single contour covering the area of both confused sources. With this in mind, it is understood the large difference of 2.4570 between our GCES system and the true IRAF reference. Therefore, our GCES contouring system has an important limitation for those cases of blended sources.

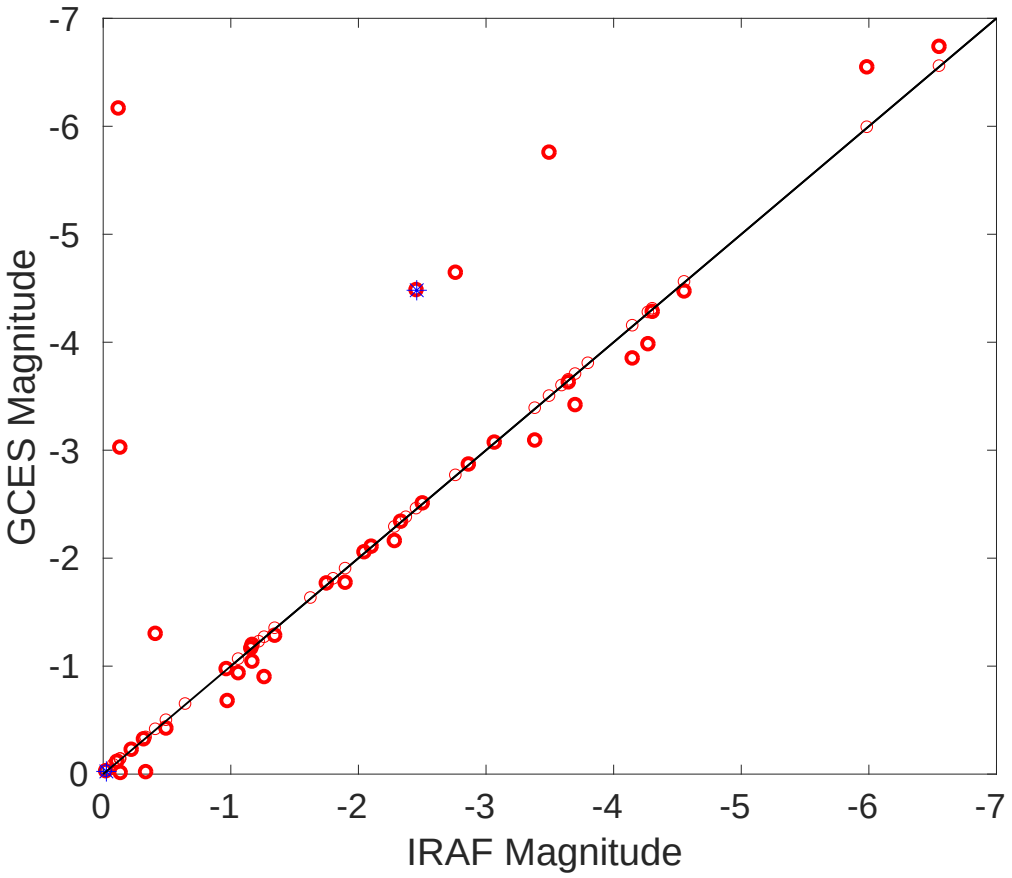

Figure 4.20: Good and Bad cases (marked with blue asterix) of GCES contours.

Figures 4.23 to 4.26 shows the outcome of running our GCES contouring algorithm with the same 50 Uniform sources as above in an image with a representative value of noise. This level of noise was chosen based on a visual comparative inspection between the synthetic images obtained here and the real COSMOS images used in Chapter 5. Here we are interested in analysing how the noise affects to the performance of our GCES contour system for the cases of isolated sources. Figure 4.23 shows the synthetic image identical to the image of Figure 4.20 but with the addition of noise. Figure 4.24 shows the magnitude values yielded by our GCES system versus the IRAF true magnitude values. Comparing this figure with Figure 4.20 we observe that the alignment with the diagonal is kept, but with a more clear dispersion in their separation from the diagonal. The reason for this difference is the noise added. Figures 4.25 and 4.26 show the GCES contours for the same two cases as in the Figures 4.21 and 4.22. We observe that for the case of isolated source, shown in Figure 4.25 the GCES contour is generally preserved, compared with the contour of the same object

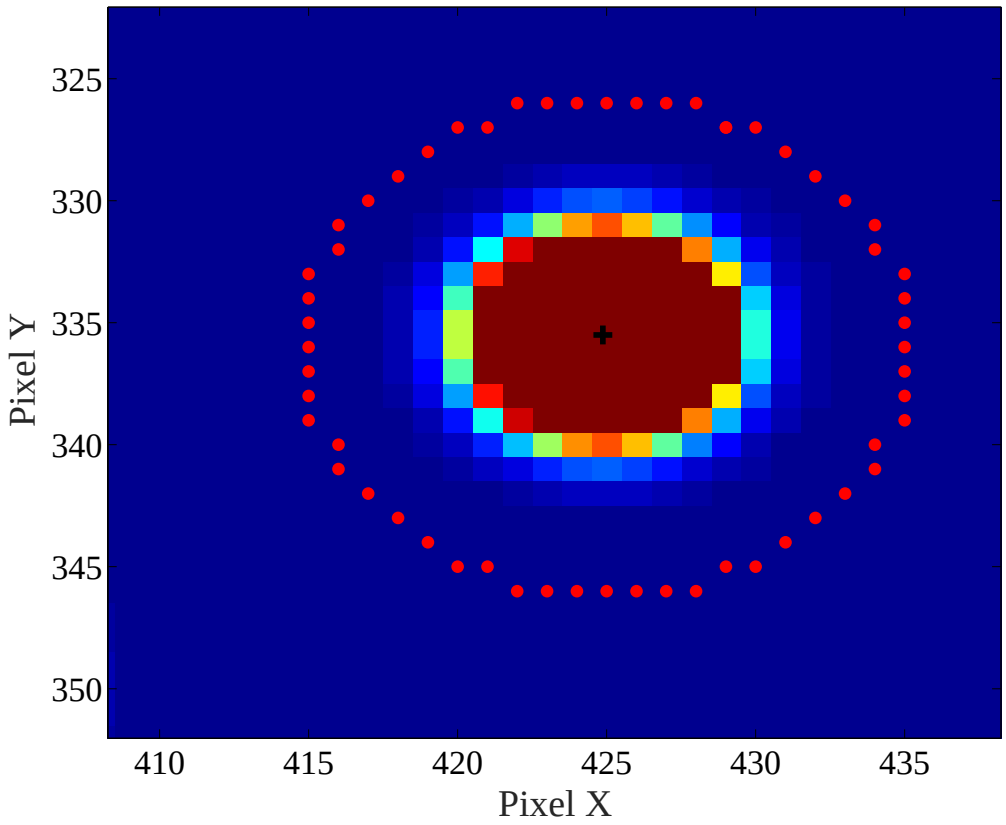

Figure 4.21: Case of small difference ($-1.3661e-04$) between the values of magnitudes of IRAF and GCES.

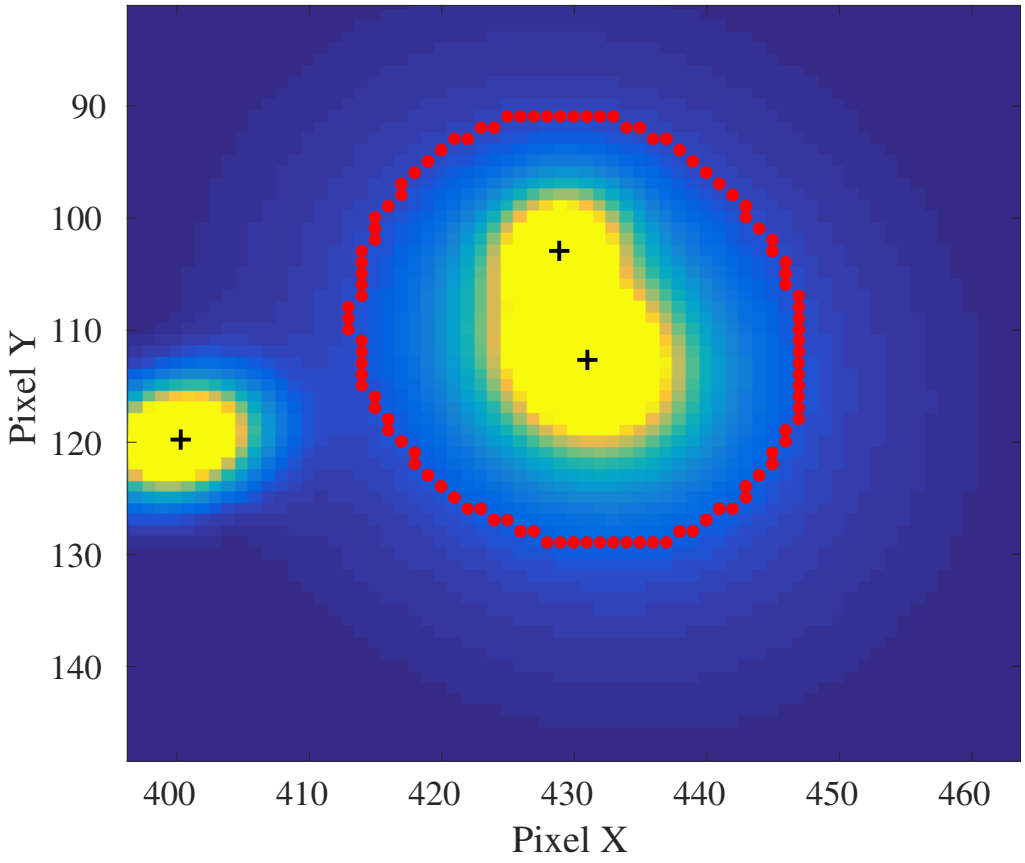

Figure 4.22: Case of large difference (2.0243) between the values of magnitudes computed by IRAF and by Active Contour.

in the image without noise, shown in Figure 4.21. This probes that the GCES contour algorithm can handle this level of noise. For the case of blendded sources and noise, shown in Figure 4.26, the GCES contour performance is clearly worse than in the absence of noise or very low level of it. In effect, the GCES contour of Figure 4.26 does not cover all the area of the two blended sources, contrary to the result shown in Figure 4.22. Therefore the limitation of our GCES contour algoritm for blended sources is stressed when we consider a realistic noise level.

Figures 4.27 to 4.30 present the behaviour of our GCES contour system on a noisy and crowded

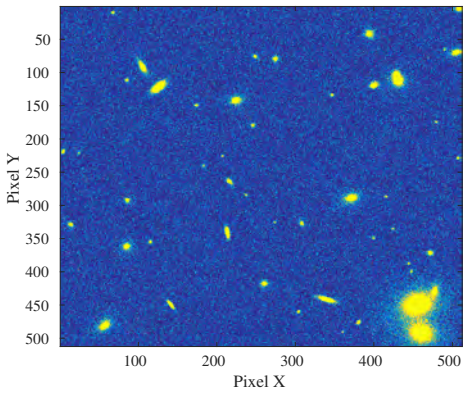

Figure 4.23: Artificial image created from a Uniform distribution with rd noise 20 and poisson noise. 50 galaxies created. The active contour are represented in white.

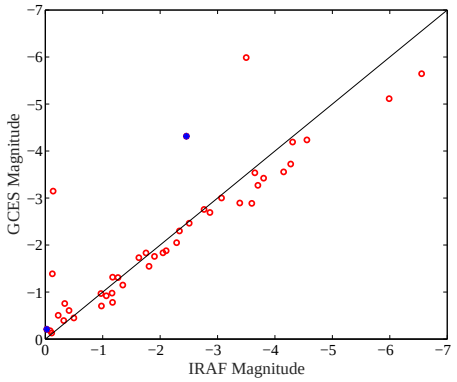

Figure 4.24: Comparison on the magnitude values yielded by IRAF and the ones from the active contour.

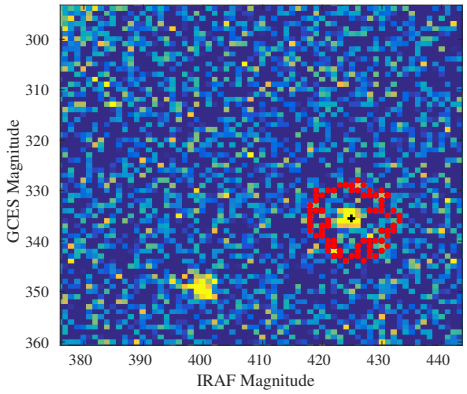

Figure 4.25: Case of small difference (0.1839) between the values of magnitudes computed by IRAF and by Active contour.

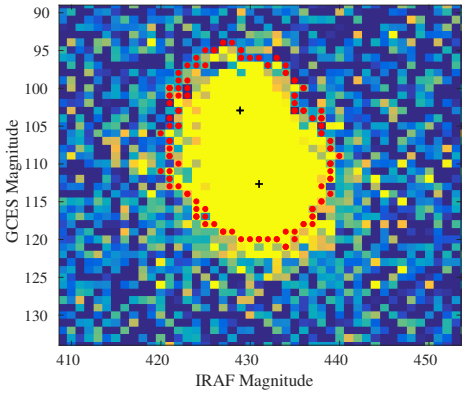

Figure 4.26: Case of considerable large difference (1.8575) between the values of magnitudes computed by IRAF and by Active Contour.

image. Here we are interested in analysing the main limitations of our system. Therefore we will stress our GCES system with a synthetic image that contain levels of noise worse than the ones typically found in a real image. The image of Figure 4.27 shows 100 synthetic noisy sources in a

Hubble crowded distribution. Figure 4.28 shows the GCES magnitude values versus the IRAF ones. From this figure we observe one group of cases located close to the diagonal, with certain level of dispersion, due to the noise and crowding of the image, and another group of cases largely out of the diagonal, corresponding to group of blended sources. The distribution of this synthetic data is Hubble and not Uniform as the two cases above. With this distribution there is an important concentration of blended sources in the center of the image, and this explains the large number of cases located largely far from the diagonal of figure 4.24. We selected two objects, one with GCES magnitudes very close to the diagonal, and another one with a GCES magnitude value clearly separated from the diagonal; these two cases are marked in blue dot in Figure 4.28. Figure 4.29 shows the contour of an isolated source for which the GCES magnitude is very close to the IRAF reference one. We observe that this contour is covering the area of the noise object. Figure 4.30 shows the GCES contour for the case of a group of noisy blended objects. A large difference in magnitude between GCES and IRAF is found. The noise and blending are the two factors affecting this large difference.

Throughout these results it is observed that the GCES system retains the capability of absorbing certain level of noise (read noise and Poisson). In effect, if we compare Figures 4.20 and 4.24 we observe that overall the result is similar. The image of Figure 4.18 has no noise and while the image of Figure 4.23 has read and Poisson noise. The contour offered by our GCES system shows, as expected, some degree of additional curvatures when confronted with an important degree of noise. The performance of our GCES system deteriorates considerably with noisy and crowding. The main reason for that can be found in the active contour termination condition. The number of iterations of the algorithm has been considered as a modifiable factor depending of the nature of the image in terms of population density and/or noise. In line with this, figures 4.27 to 4.30 present the result of our GCES system for 100 Hubble galaxies in an image with read and Poisson noise and a poor gain of 1. Figure 4.28 shows an important fraction of sources with very large divergences in magnitude values compared to the ones produced by IRAF. These large differences correspond mainly to areas in which the active contour is gathering more than one source because there is a high population density in that location.

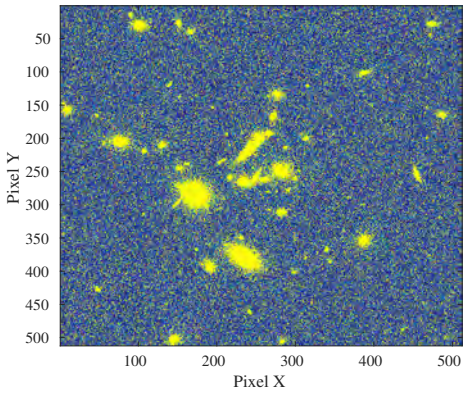

Figure 4.27: Artificial image created from a Hubble distribution with gain 1 readnoise 10 and poisson noise. 100 galaxies created. The active contour are represented in white.

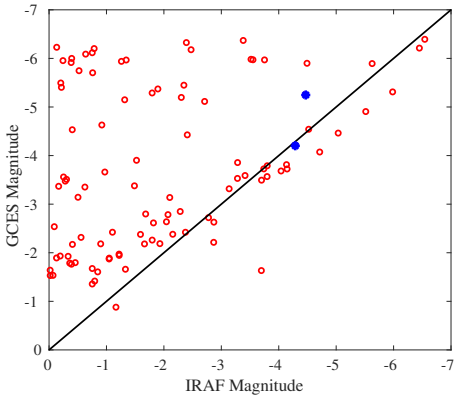

Figure 4.28: Comparison on the magnitude values yielded by IRAF and the ones from the active contour.

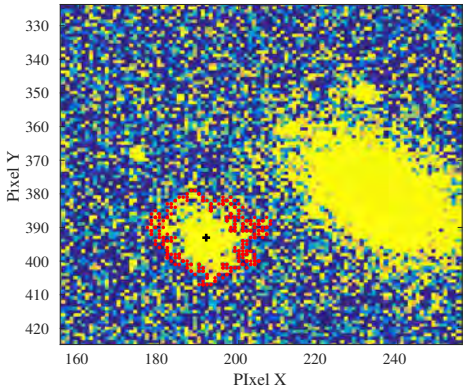

Figure 4.29: Case of small difference (0.1348) between the values of magnitudes computed by IRAF and by Active contour.

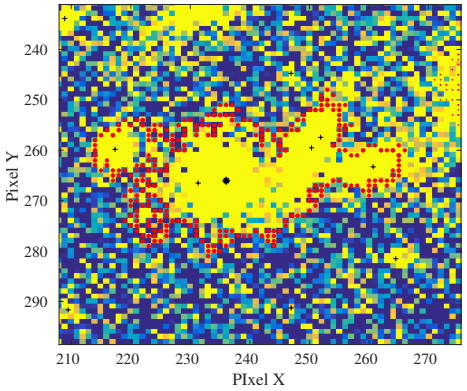

Figure 4.30: Case of considerable large difference (0.7186) between the values of magnitudes computed by IRAF and by Active Contour.

We created a set of synthetic images and SExtractor catalogues in order to observe and assess the behaviour of our contour algorithms compared to the values obtained from SExtractor with no configuration tunning. The outcomes from this exercise is shown in Figures 4.31 to 4.34. However, a validation assessment cannot be extracted from these results.
Figures 4.31 and 4.32 show the results obtained for the validation of the active contour algorithm on 24 and 28 sources created artificially using IRAF and following the Uniform and Hubble spatial distribution respectively. It can be observed that the source contours yielded by the active contour algorithm cover all the pixels belonging to the sources, according to a careful visual inspection on isolated sources. It can also be observed that the Kron ellipses obtained from the traditional SExtractor tool are leaving out of the contour pixels which clearly belong to the source under study. This happens specially with very extended sources. Therefore, we can conclude that in general for the cases of isolated sources and a signal to noise ratio (S/N) of 3 the active contour algorithm reaches a better accuracy than the traditional Kron ellipses in determining the boundary of the sources, without requiring aperture corrections. The immediate consequence of this is reflected in the surface brightness computation which yields a larger apparent magnitude values for the Kron ellipses than for the active contours.
Regarding the Matlab contour function, from the results with artificial data shown here we can conclude that this approach does not retain the same level of accuracy in the identification of the brightness gradient, as it can be achieved with the active contour approach. Furthermore, when comparing the Kron ellipses and the Matlab contour function for artificial data – see 4.33 and 4.34, the Kron ellipses yield in general better accuracy in the contour determination than the Matlab contour function but still comparable. Conversely, in the case of extended sources with a clear differentiation between foreground and background, the Matlab contour function reaches the contour effectively and with less computational cost than the active contour.
Figures 4.33b and 4.34b shows that the magnitudes obtained from the Matlab contour function are above the reference line, while figures 4.31b, 4.32b show the active contour magnitudes below the reference line. This is due to the main different behaviour between the two contouring approaches explored in this research. The Matlab contour function operates on the basis of isolines, whereas our modified active contour considers the problem of maximum gradient of brightness as a local problem of each source; therefore, the modified active contour algorithm computes a more generous contour for the sources than the MATLAB contour and that the Kron ellipses define here. Figure 4.33d shows a superior performance of the Kron ellipses compared to the MATLAB contour in front of blended objects. Based on the validation analysis with IRAF synthetic data done for the GCES contour and blended objects, the same superior performance for blended objects of Kron ellipses is expected over the GCES contours.

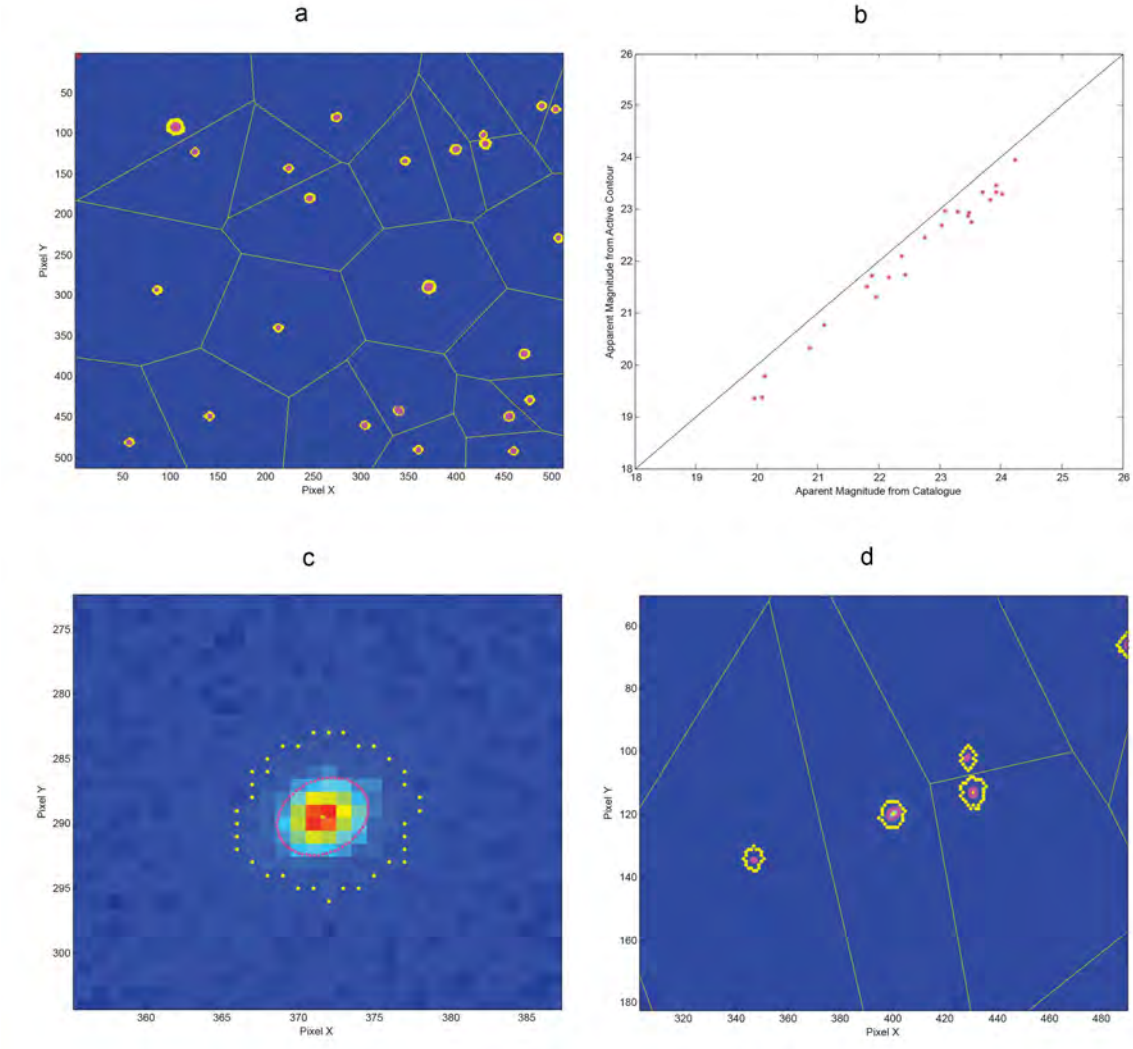

Figure 4.31: 24 IRAF artificial sources following a Uniform distribution, with Poisson noise and $S/N = 3$. Fig a. Voronoi tessellation and active contour (yellow) and Kron ellipses (magenta). Fig b. apparent magnitude from SExtractor catalogue (x axis) and from active contour algorithm (y axis). Fig c. zoom showing Kron and active contour on the same source. Fig d. zoom showing Voronoi cells with isolated sources.

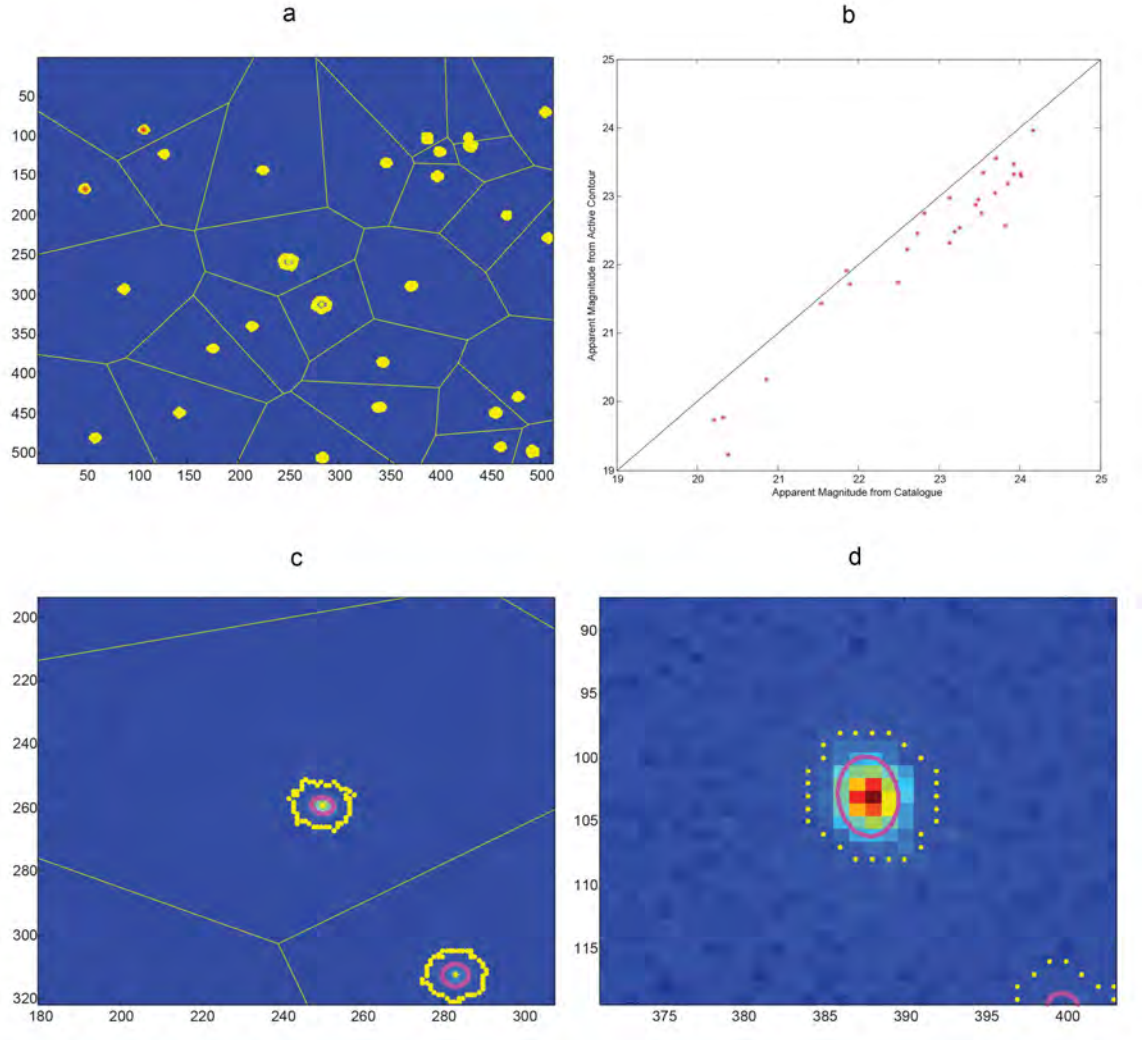

Figure 4.32: 28 IRAF artificial sources following a Hubble distribution, with Poisson noise and $S/N = 3$. Fig a. Voronoi tessellation and active contour (yellow). Fig b. apparent magnitude from SExtractor catalogue (x axis) and from active contour algorithm (y axis). Fig c. zoom showing Voronoi cells with isolated sources in contiguous Voronoi cells. Fig d. zoom showing Kron and active contour on the same source.

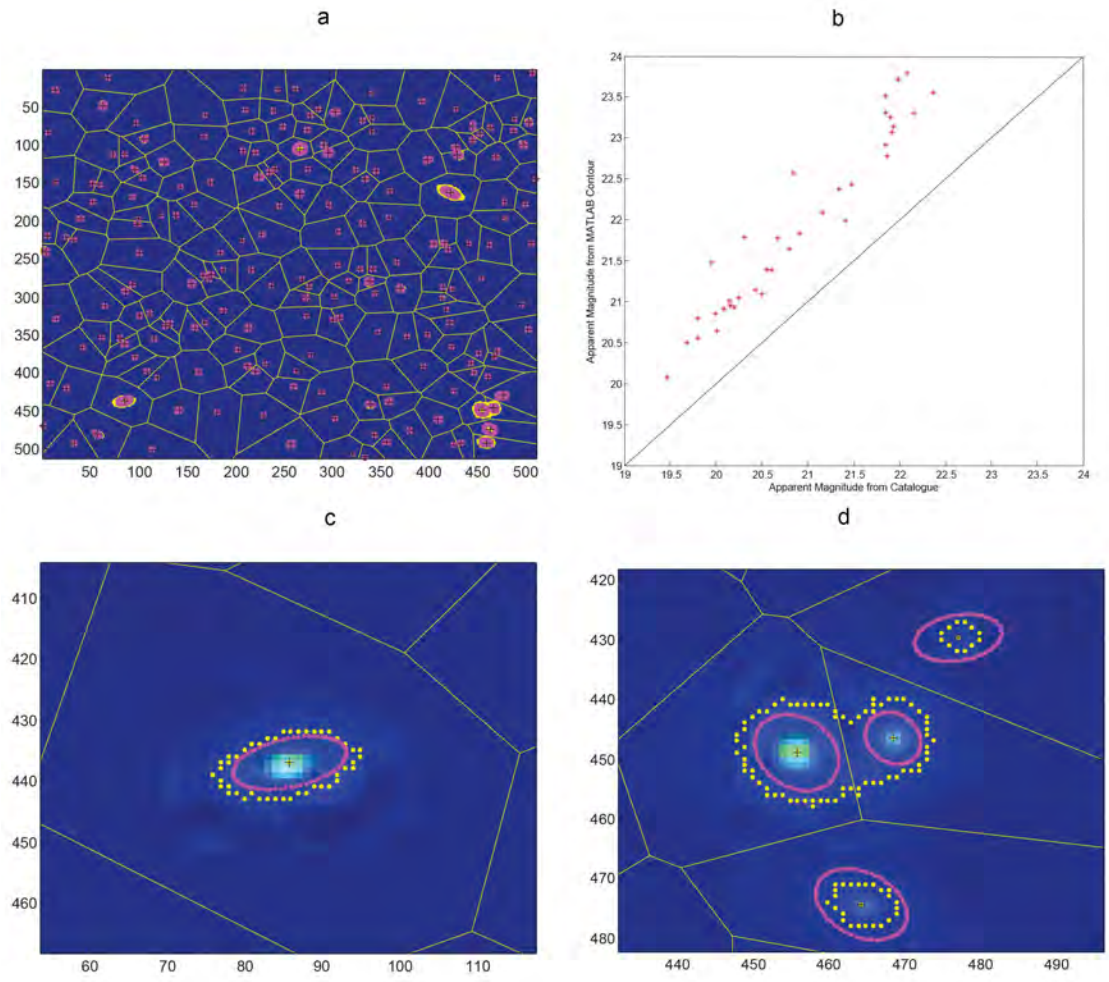

Figure 4.33: 194 IRAF artificial sources following uniform distribution, with Poisson noise and $S/N = 3$. Fig a. Voronoi tessellation and MATLAB contour (yellow) and Kron SExtractor ellipses (magenta). Fig b. Apparent magnitude from SExtractor catalogue (x axis) and from MATLAB contour algorithm (y axis). Fig c. zoom showing the MATLAB contour and the Kron ellipse on one isolated source. Fig d. Another zoom showing a few sources including blended ones and the MATLAB contours yielded with this challenging configuration.

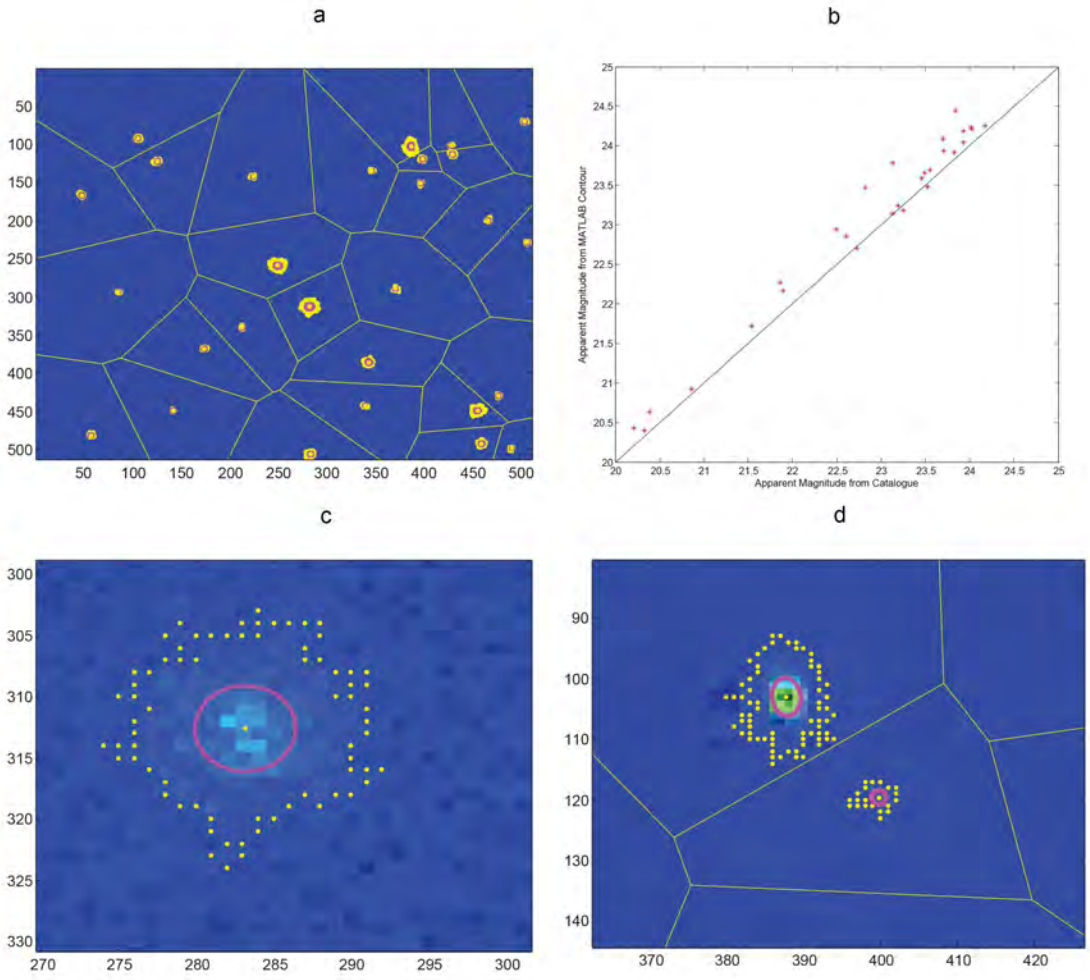

Figure 4.34: 28 IRAF artificial sources following the Hubble distribution, with Poisson noise and $S/N = 3$. Fig a. Voronoi tessellation and Active contour (yellow) and Kron SExtractor ellipses (magenta). Fig b. Apparent magnitude from SExtractor catalogue (x axis) and from MATLAB contour algorithm (y axis). Fig c. zoom showing the active contour and the Kron ellipse on one isolated source. Fig d. Another zoom showing a couple of isolated sources in contiguous Voronoi cells and the active and Kron contours.

4.6 Source Labelling Validation

The validation of the methodology implemented in the automatic classification of isolated versus non-isolated galaxies in FITs images was carried out using artificial images and artificial catalogues, as shown in Figures 4.35 to 4.38.
It is important to note here that this rule-based system for the labelling of sources is a deterministic problem and therefore, once the rules are established and validated for one case, the behaviour of the algorithm will be completely fixed. The only limitation found in this part of the GCES is in terms of computational cost. The Voronoi tessellation and the exploratory rule algorithm per source and per cell of the entire image can reach the limit of MATLAB. The solution to this limitation consists in considering appropriate windows of study which is commensurable to the computational capability of the system used.
From the results of Figure 4.35 to Figure 4.38 we can observe that the isolated, partially contaminated and contaminating sources are correctly marked with green, yellow and red in their respective contours as expected from our rule based system. It is important to note here that for those cases for which the Voronoi cell contains at least one vertex in the infinity, the rule system may lead to unreliable results, and therefore we rule out those cases. The sources for which the Voronoi cell has at least one vertex in the infinity are flagged with a black asterix in their pixel coordinates and they are not considered as part of our GCES labelling system. In terms of application of our GCES labelling system with real data, this scenario typically corresponds to those cases of sources in the fontier of the image tiles. However, these sources are replicated into the neighbour tiles such that their position is far from the frontier in the new tile. Therefore the GCES labelling system is normally not impacted by this scenario when used with real data.
One important consideration here is the fact that our GCES rule based system peforms a preliminary identification of the condition of contamination of each source, primarily based on the level of contamination at its Voronoi cell. The result of this needs to be re-assessed in view of the possible inter-contour contamination. In effect, as a preliminary result from our rule based system we could have a scenario in which a source is fully contained in its Voronoi cell and another neighbour source is contaminating the Voronoi cell of the first source but not the contour of it. In that case, the label remains isolated for the first source.
Another relevant aspect to discuss here is the notion of multi-band labelling of sources. The practical implementation of our GCES system is focussed on multi-band deep surveys. In this scenario it seems logic to compile the outcomes of the GCES labelling per source and per band into a multi-band concept of labelling. Following this idea, our GCES labelling rule based system would only consider isolated sources those that are labelled as isolated in all the bands of the survey under consideration. Similarly it happen for the labels of partially or dubious contamination and contamination. The two considerations above are not part of the scope of this research but they are implicitly included in one of the future lines of work identified in Chapter 6.

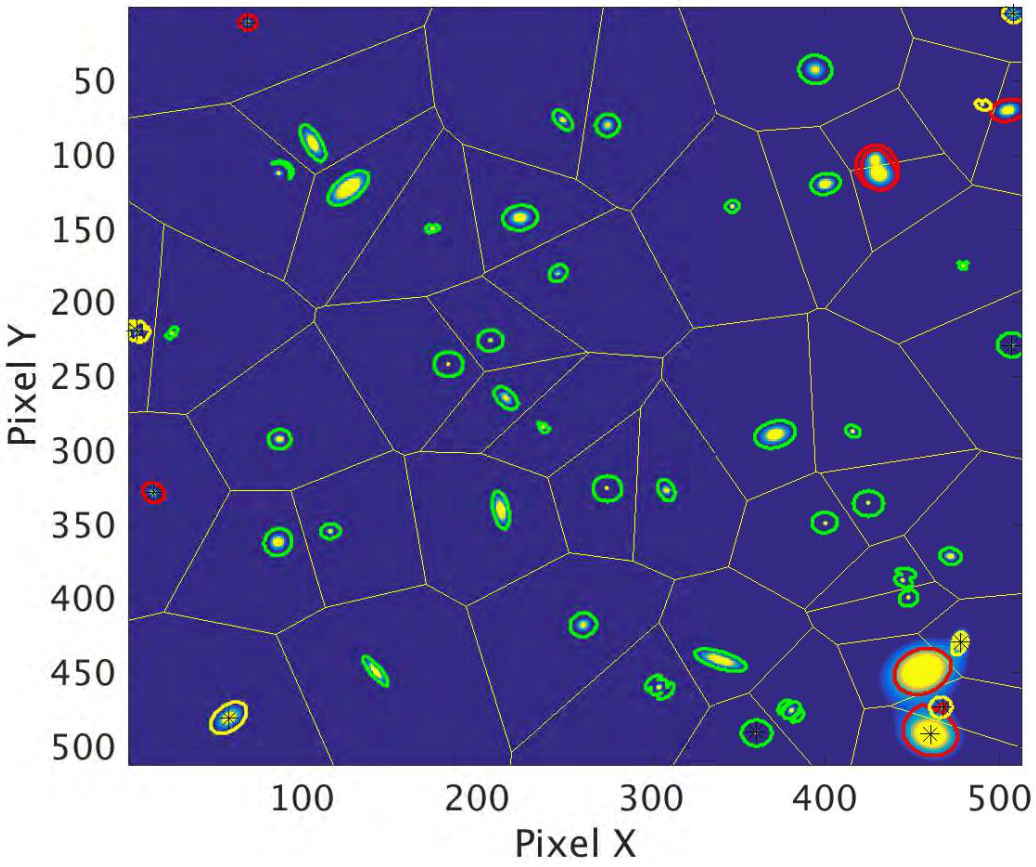

Figure 4.35: Outcome of Source Labelling algorithm for 50 artificial Uniform galaxies with gain 10. Isolated sources are labelled with green colour, partially contaminated sources are represented in yellow and contaminated sources in red.

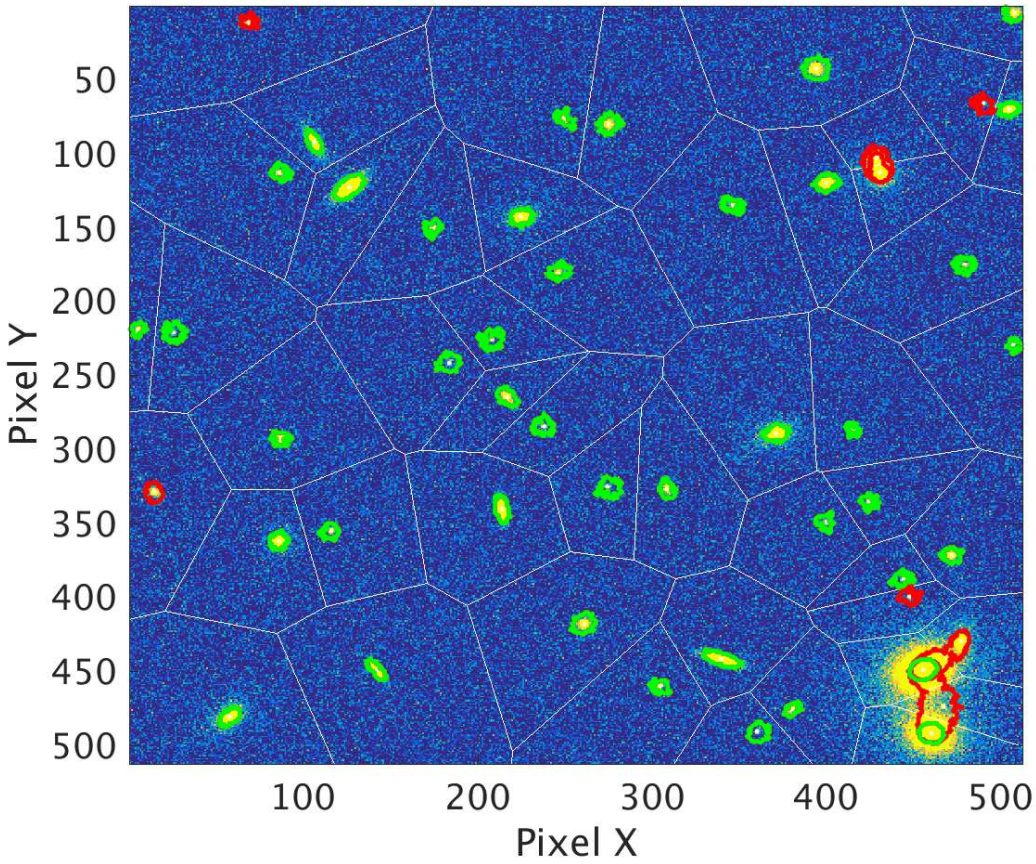

Figure 4.36: Outcome of the Source Labelling algorithm for 50 artificial galaxies distributed uniformly across the image with gain 10, read-noise 20 and Poisson noise. Isolated sources are labelled with green colour, partially contaminated sources are represented in yellow and contaminated sources in red.

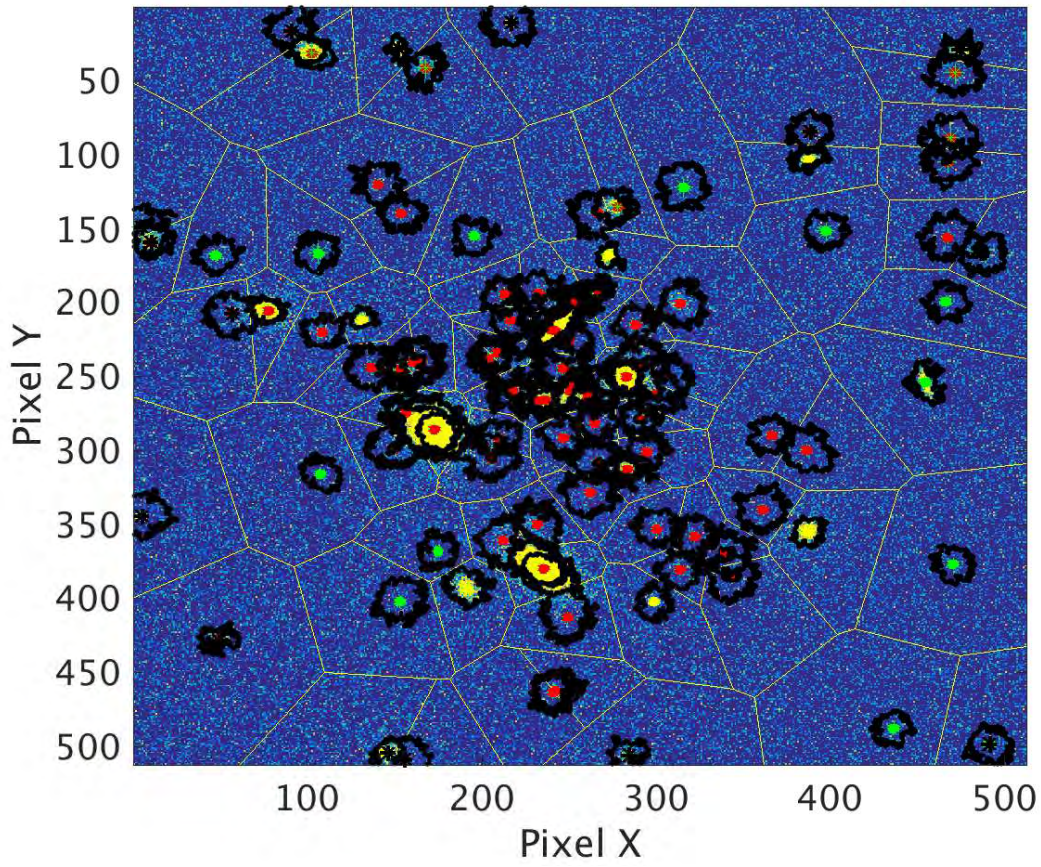

Figure 4.37: Outcome of Source Labelling algorithm for 100 artificial galaxies following the Hubble spatial distribution with gain 1 read-noise 10 and poisson noise. Isolated sources are labelled with green stars, partially contaminated sources are represented in yellow and contaminated sources in red.

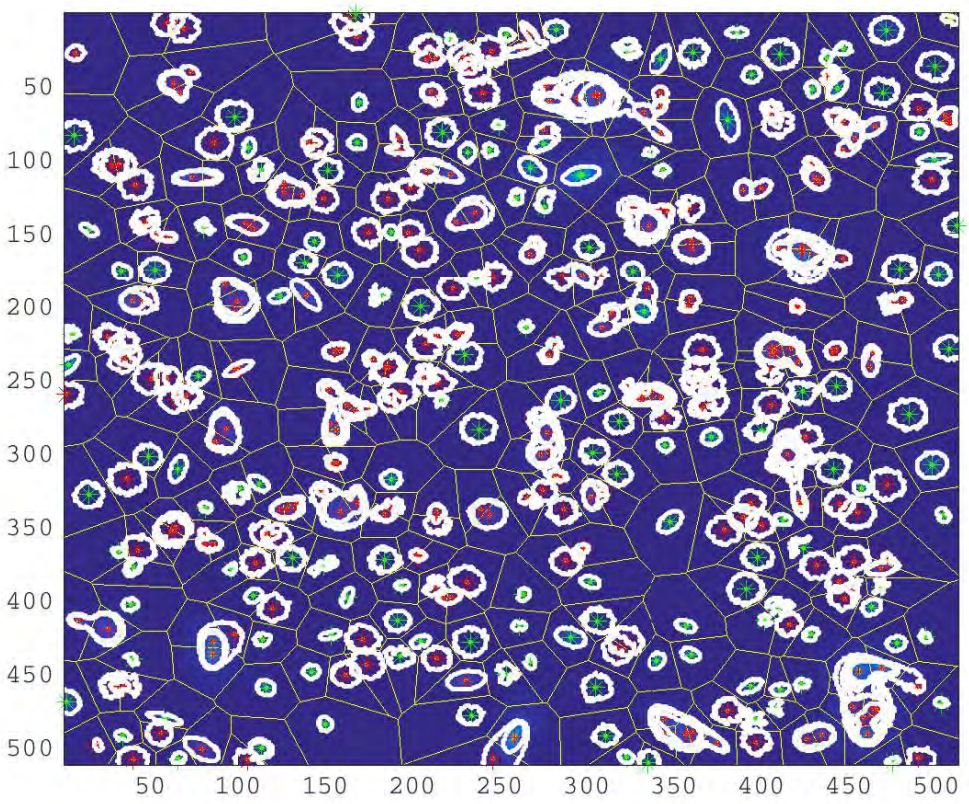

Figure 4.38: Outcome of Source Labelling algorithm for 425 artificial galaxies uniformly distributed across the image with gain 10, read-noise 10 and poisson noise. Isolated sources are labelled with green stars, partially contaminated sources are represented in yellow and contaminated sources in red.

Chapter 5

GCES Results with COSMOS Data

5.1 Introduction

In this chapter we validate our expert system (GCES) with real COSMOS data sets. We present the results obtained from the different use cases and, based on it, we provide an assessment of the contributions and limitations.
As indicated in [Capak et al., 2007], the advances in Astronomy are often driven by improved accuracy and precision along with acquisition of more data. Photometric redshifts are susceptible to systematics in all bands and this increases the calibration requirements. This chapter is based on the COSMOS data set which is obtained from various space and on-ground instruments, each of them with different measurement set-ups and data acquisition methodologies, including software reduction and configuration. The data used in our research includes the implementation of the data reduction steps as described in [Capak et al., 2007], which ensures a high level of photometric accuracy.
This chapter introduces in Section 5.2 the overall context of the data and model, including a list of assumptions made. [Capak et al., 2007], [Ilbert et al., 2009] and [Laigle et al., 2016] contain further detailed descriptions of the COSMOS Survey. Section 5.3 offers a detailed description of the results in each of the main modules, as well as results from the use of the full GCES expert system. Then Section 5.4 presents a critical assessment of the results. The main goal in this Section is to help the reader in identifying the main contributions, a comparative analysis with traditional methods and the main limitations.

5.2 The COSMOS Survey

COSMOS imaging uses a simple grid of sub-images for all data products. The area is divided into 144 sections named tiles of $10' \times 10'$, and each section is covered by an image of size 4096×4096 pixels with a pixel scale of $0.15''$. Adjacent tiles overlap each other by $14.4''$ on all sides. As a result, we can analyse the vast majority of objects on a single image. The pixel scale was chosen to be an integer multiple of the $0.05''$ scale used for the HST ACS images. The units to which we scaled all images and noise maps are nanoJanskys per pixel, which corresponds to zero point magnitude of 31.4. Figure 5.2 shows the images obtained using MATLAB for Tile 78. The

132	133	134	135	136	137	138	139	140	141	142	143
120	121	122	123	124	125	126	127	128	129	130	131
108	109	110	111	112	113	114	115	116	117	118	119
096	097	098	099	100	101	102	103	104	105	106	107
084	085	086	087	088	089	090	091	092	093	094	095
072	073	074	075	076	077	078	079	080	081	082	083
060	061	062	063	064	065	066	067	068	069	070	071
048	049	050	051	052	053	054	055	056	057	058	059
036	037	038	039	040	041	042	043	044	045	046	047
024	025	026	027	028	029	030	031	032	033	034	035
012	013	014	015	016	017	018	019	020	021	022	023
000	001	002	003	004	005	006	007	008	009	010	011

Figure 5.1: COSMOS tile sub images: Figure taken from IRSA website

results and practical contribution of this research were focused on this tile. We selected this Tile for two main reasons: it has a centred position in the overall mosaic, see Figure 5.1, and the total number of sources is in the average considering corresponding values for the rest of the tiles.

The information of this Section has been extracted from [Capak et al., 2007] in order to provide the reader with a clear picture of the characterized input to the expert system of this research, and also to aid the reader in fully understanding some of the assumptions enumerated above. It also identifies which steps were done manually (involving a direct human operator task) in the production of the COSMOS catalogue release of 2008 and lately 2015 and of the images published in the NASA-IRSA Web site.

We used the two photometric catalogues, COSMOS Intermediate and Broad Band Photometry Catalogue 2008, alias **cosmos phot 2008** and COSMOS Photometric Redshift Catalogue Fall 2008 alias **cosmos zphot mag251985**, and the corresponding FITS images from the COSMOS survey obtained via the NASA/IPAC Infra-red Science Archive web site,[1] as a good candidate for real calibrated data. In addition to that, Cosmos Morphology Catalogue from 2005 was the only one found for which Kron ellipses parameters were available. However due to the date of release of that catalogue compared to the date of the photometric catalogues used, the comparisons were performed as a preliminary trial to visually inspect the differences between SExtractor Kron ellipses and the active contours.

This photometric catalogue has a well-defined SED template library that contains 31 templates, including shapes such as elliptical and spiral, and a range of redshift up to 6. The following bands and associated transmission filters were considered: CFHT u*, Subaru B, Subaru V, Subaru g+, Subaru r+, Subaru i+, Subaru z+, CFHT i', UKIRT J, and CFHT K, with an effective wavelength range from 3798 to 21460 Angstroms.

[Capak et al., 2007] provides a detailed description of the catalogue. All photometry presented here corresponds to the AB magnitude system. The catalogue content is composed of PSF-matched 3" aperture photometry and error (or measurement deviation) for all Subaru and CFHT bands, along with UKIRT bands and other bands not used in this research.

Objects with no detection are assigned a magnitude of 99. Objects with no measurement due to lack of coverage, saturation or other defects are assigned a magnitude and error of -99. Of special relevance to our research here are the flag columns at the end of the catalogue. The cleanest catalogue will have all flags set to zero. The photometry flags indicate the area of the photometry aperture in square arc second which is in a masked region. Objects with a deblending flag set to 1 indicate that they could be spurious and in principle, we should not use them for statistical studies. However, the photometry for real sources with a deblending flag set to 1 is good if they fall outside of the masked region. Finally, a flag indicating that the object may be a star instead of a galaxy is included. This flag is, however, a qualitative assessment, so again we observe here a possible **manual step**, whereas an automatic one in terms of inferring the probability of a star or galaxy could provide a more effective and quantitative assessment. A line of work in this direction is described in [Fadely et al., 2012]. Therefore, if we consider a further thorough check of these cases, it might be that sources initially flagged as spurious are not finally considered as such. A refinement of the initial catalogue in terms of the quality of its data is one of the objectives of the research presented here.

The approach followed in this research consisted of ignoring the de blending and star flags when cross-correlating with multi-wavelength detections. Conversely, this research uses the quality flags to identify what will be designated as "grey areas" also named "dubious area" of the catalogue, by which it is meant catalogue information of low reliability. A study of the implementation of the complete GCES in this "grey area" was conducted, aiming at refining the quality of the catalogue in that grey area and effectively extracting potentially new useful information.

In addition to the above, for deep surveys, the ability to detect objects (completeness) and separate superimposed objects (confusion) are very important and often the magnitude of their measurement

[1] http://irsa.ipac.caltech.edu/cgi-bin/Gator/nph-scan?projshort=COSMOS

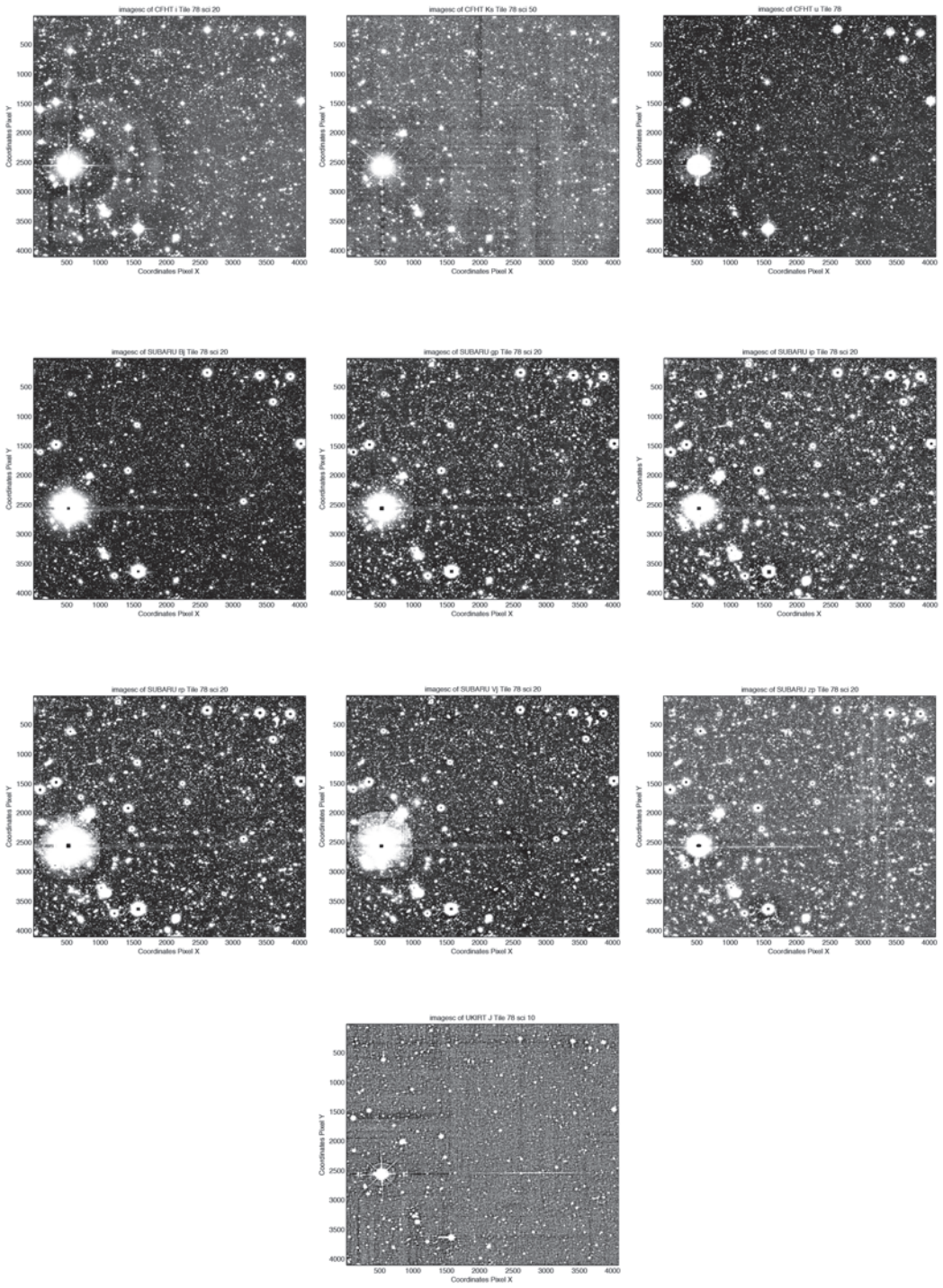

Figure 5.2: COSMOS images from the IRSA website.

is larger than the photon noise. These quantities are difficult to quantify because they are sensitive to both the data quality and the software used. In this sense, when estimating the total magnitude for each survey, efforts are best used for minimising the effects of aperture size. Our approach of computing surface brightness of the sources and identifying the isolated from non-isolated ones had to face the challenge of identifying sources in crowded areas of the sky. These two aspects are possibly the more subjective ones in terms of manual inspection of the image and at the same time they can also be the most challenging ones in terms of the alternative automatic step proposed in our pipeline. A further future line of work in this area is envisaged, mainly to address the problem of automatic determination of boundaries of blended sources in crowded and noisy areas of the image. The following assumptions and considerations, derived from [Capak et al., 2007], were taken into account when using COSMOS real data:

- When using the IRSA image map for searching, the pixel resolution of an all-sky image of COSMOS data is 0.5 degrees.
- At apparent magnitudes fainter than 25, the catalogue begins to be incomplete and to have more spurious detections and the photometric redshifts begins to behave poorly.
- Cosmic ray events are not considered to be part of the images; they are removed by detecting sharp edges in images, and every frame was visually inspected to remove internal reflections, satellites, asteroids and other false objects before the formal release of the data used.
- The photometry taken from the catalogue is initially assumed to have no specific corrections applied. The GCES expert system took the corrections recommended in [Capak et al., 2007] into consideration.
- The uncertainty declared in the Subaru B filter profile and its expected impact has not been taken into account until a more conclusive outcome of that investigation is available. However, the GCES took this aspect into consideration, should the implementation of this uncertainty become necessary.
- The photometric offset from [Capak et al., 2007] was not applied to the catalogue used in this research, since it cannot be verified using external calibration sources at this time.
- It is assumed that the input data was properly calibrated, that is to say, that the fulfilment of calibration requirements considered in [Capak et al., 2007] (e.g. adequate spatial variations in photometry and astrometry, adequate night sky subtraction process, adequate correction for residual scattered light, consideration of mechanical and optical constraints, etc.) was assumed as part of any input data to our GCES, and therefore the implementation of those calibration requirements are currently out of the scope of this research.
- The COSMOS photometric catalogues have undergone a series of manual steps briefly enumerated in [Capak et al., 2007], such as visual inspection to remove artefacts, cosmic rays, etc. However none of the manual steps indicated in the bibliography of the COSMOS survey release was considered for the development and implementation of our GCES. It is therefore assumed that the use of our GCES does not need any expert knowledge.
- Defects were rejected in the CFHT reduction pipeline by performing a manual step: images with a seeing larger than $1.3''$ in the $i*$ band and $1.4''$ in the $u*$ band were rejected.
- A factor of 2 difference in seeing corresponds to a 0.3 mag reduction in measurement sensitivity, but a 0.75 mag reduction in peak surface brightness and a factor of 4 increase in de-blending.
- The zero-point corrections given in Table 13 of [Capak et al., 2007] are not included in the published COSMOS catalogue of 2008 release.

- The zero-point corrections have an rms amplitude of ±0.06 mag, which is slightly larger than the expected error. However, the zero-points calculated as per [Capak et al., 2007] disagree with the ACS F814W photometry by −0.118 mag and fail to produce photometric redshifts free from systematic errors.
- With typical overheads of 15 minutes per standard, it is extremely difficult to obtain a sufficient number of standard star on Suprime-Cam. With three to five standards per band, the COSMOS standard-star calibrations are accurate to ±0.05 mag. These offsets are larger than desired for accurate photometric redshifts.
- The full catalogue should be used with caution. In particular, the completeness and the number of spurious sources will vary as a function of position due to differences in the rms background noise.
- According to [Capak et al., 2007], the COSMOS data have the extremely high level of photometric consistency necessary for scientific pursuits such as large-scale structure studies.
- The expected zero-point variations are $\leq 1\%$ across the field; this level of photometry was achieved by constructing flat fields directly from object fluxes rather than using sky or dome flats.
- The tuples of the various COSMOS catalogue releases are considered astrometric matches, i.e. the sources of each tuples were selected based on their astrometric proximity.

5.3 Results

This section presents a summary of the compilation of the results for a representative area of COSMOS data set; Section 5.3.1 presents the results from implementing the photometric bayesian inference with COSMOS Data Set in the cross matching of photometric multi-band tuples; Section 5.3.2 presents the results from determining the active contours with COSMOS Data Set as part of the surface brightness computation and connected AB magnitude values, Section 5.3.3 presents the results of COSMOS sources labelling in categories of isolated, contaminated and partially contaminated sources, as result of the Voronoi tessellation and the implemented rule based system. Section 5.3.4 presents indicative results from using the full GCES in a feedback configuration, as shown in Figure5.72. With this we can present representative results GCES capabilities when used in full.

5.3.1 Photometric Cross-Matching

We ran the photometric cross-matching algorithm on a subset of representative sources (around 2500 sources) with high and low quality flags. As expected we observed good correspondence between high/low values of photometric Bayes factor and good/bad SED fitting, respectively. We consider good SED fitting when there is at least one SED from the grid such that each of its elements (model) is found within the error bar of each element of the tuple (data) for which we calculate the SED fit. Conversely we consider a bad SED fitting when at least one element of the model is found outside the error bar of its corresponding element of the data.
In addition to that, we have executed the photometric cross matching in the full COSMOS catalogue with low quality flags .
In line with the relevant bibliography on Bayesian Inference, we consider the following photometric Bayes Factor boundary to determine true match, spurious match or uncertainty. Therefore we consider the following cases:

- Low value of photometric Bayes factor (value less than 1 in logarithmic scale) and poor SED fitting: spurious match.
- High value of photometric Bayes factor (a value above 1 in logaritmic scale) and good SED fitting: true match,
- Value of photometric Bayes factor around 1 (in logaritmic scale): cannot be determined if it is a true match or a spurious match.

The figures in this section present the result of computing the photometric Bayes Factor for the above priors described in Chapter 3 and for a total of 2656 tuples of the 10 bands listed above. Some of these figures represent the histograms associated with the distributions of matches and non-matches of a COSMOS photometric catalogue. In all these cases, the model - SED templates convolved with the relevant filters - is softened by a Smooth Factor of 0.03 included in the error as defined in previous sections.
Figure 5.3 shows four cases of SED fitting for different photometric matching results. A correspondence between very low photometric Bayes factor and very bad or non existing SED fitting can be observed. Conversely, high photometric Bayes Factor normally yield very good SED fitting. This means for example that among COSMOS catalogues tuples with low quality flags, those with low values of photometric Bayes factor and poor or non existing SED fitting are suspiscious of not being actually the same source.

A named l-way approach has been executed on a subset of COSMOS data composed of the six bands of SUBARU ((U,B,V,g,r,i) split into two subsets (the three redder and the three bluer sources of the 6-tuple), so it is referred to in the figures as the 2-way approach, whereas the n-way approach has taken the 10 bands from the COSMOS photometric catalogue used based on the assumption that the flux measured in each band is independent of the others (it is referred in some figures as "8-way" approach).
In addition to the validation step with synthetic data described in Chapter 4, different values of the model Smooth Factor η_f were explored. Figure 5.4 and 5.5 show the results of computing a cross-matching Bayes Factor for a n-way approach (named in the figures "8-way").
Regarding the influence of the smooth factor we observe that in general the larger the smooth factor is, the histograms shrinks, meaning that at the limit we obtain no evidence from data. Conversely, for smaller smooth factor, the range is wider and the data meaning is relevant. An interesting exercise was to find out the value of the smooth factor for which we obtain the highest peak in the histogram. The value of 0.03 for the smooth factor seemed to reach a good compromise.

The following figures show relevant aspects of of the implementation of the cross- matching problem with different priors in the COSMOS photometric catalogues. Aspects relevant to the influence and distribution of error and smoothing factor for the various priors considered are also included in this set of graphics.
To test how good a discriminator the photometric Bayes Factor is on the real COSMOS tuples, we could select a photometric Bayes Factor threshold and calculate the fraction of good or true matches found and the fraction of false positives. We plot this point on an x-y plot. If we repeat this for all possible photometric Bayes Factor thresholds values, we get a curve which allows us to evaluate the method. We performed this exercise also with COSMOS shuffled tuples which represent the case of real no match tuples. In practice we have selected around 100 photometric Bayes Factor thresholds and computed this ratio for the three priors considered in the research. It is important to note there that this test is based on the assumption that the COSMOS tuples are initially set up as astrometric matches, therefore they are all considered positive matches (those tuples with low quality flags were not part of this test).

Therefore, a fundamental question to be answered at this stage is how well the Bayes Factor discriminates between matches and no-matches. Figures 5.6 to 5.9 show the ratio of true and false positive match for a defined range of Bayes Factor thresholds. These results offer an interesting

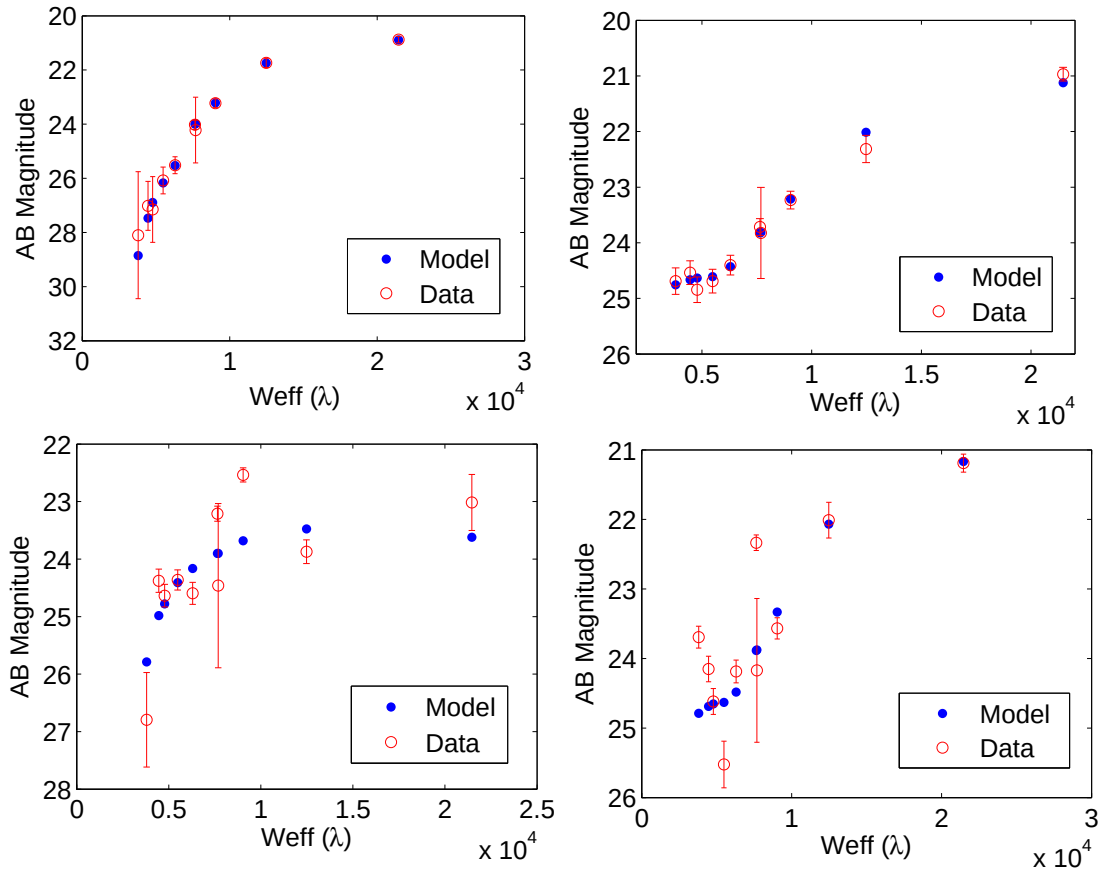

Figure 5.3: Top panel left. Best SED fit for a true matched tuple of a COSMOS galaxy of type $Ell3\_A\_0$, redshift of 1.95, colour index $B-V$ of 0.9377. Top panel right. Best SED fit for a true matched tuple of a COSMOS galaxy of type Sc_A_2, redshift of 3.1, colour index $B-V$ of -0.1523. Bottom panel left. Best SED fit for a shuffled tuple of a COSMOS galaxy of type $Ell3\_A\_0$, redshift of 0.15, colour index $B-V$ of 0.0137. Bottom panel right. Best SED fit for a shuffled tuple of a COSMOS galaxy of type Sc_A_2, redshift of 3.15, colour index $B-V$ of -1.3733. This figure shows that our models reasonably fit real COSMOS galaxies. This is not true for random associations.

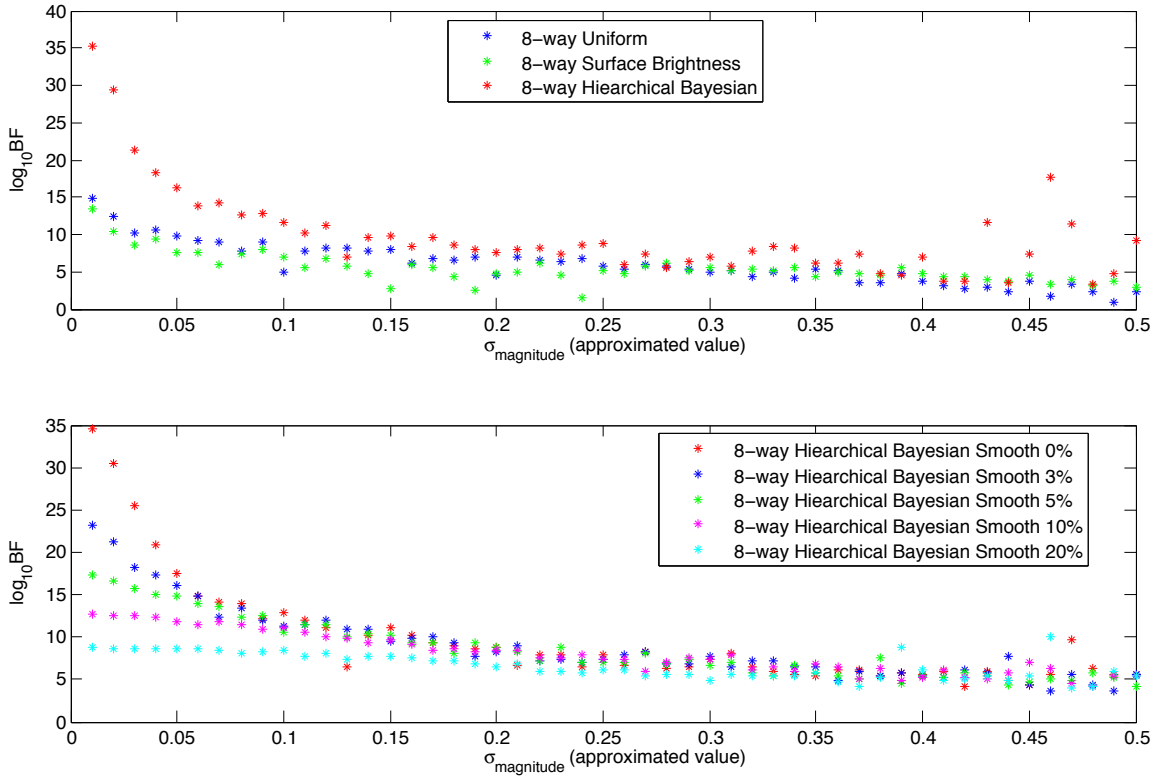

Figure 5.4: Influence of error and Smoothing Factor for the different priors in an 8-way cross-matching (from [Marquez et al., 2014]).

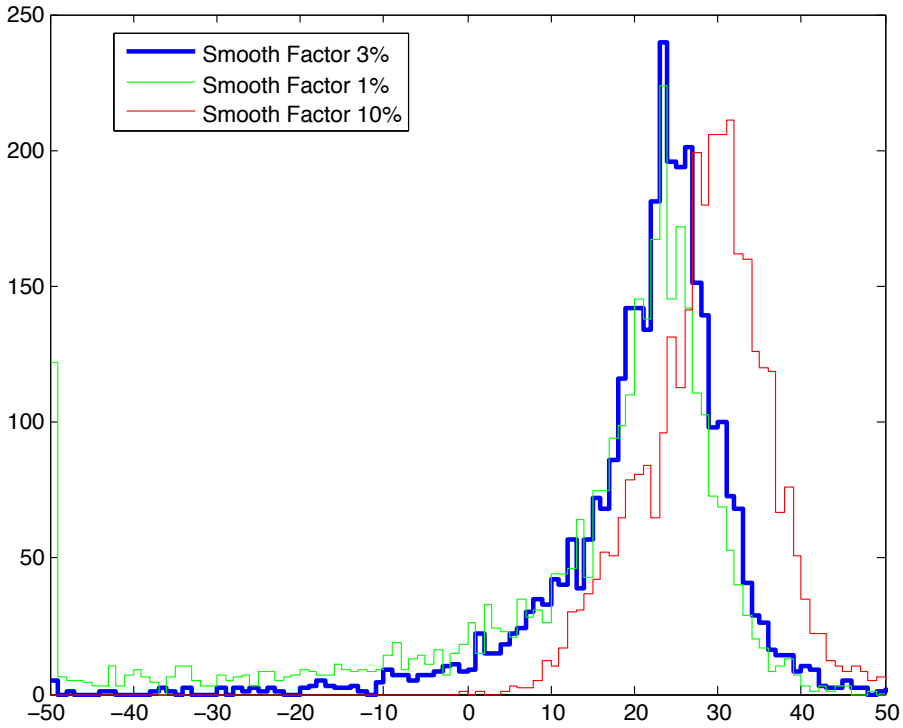

Figure 5.5: Distribution of error and Smooth Factor for the different priors in an 8-way cross-matching (from [Marquez et al., 2014]).

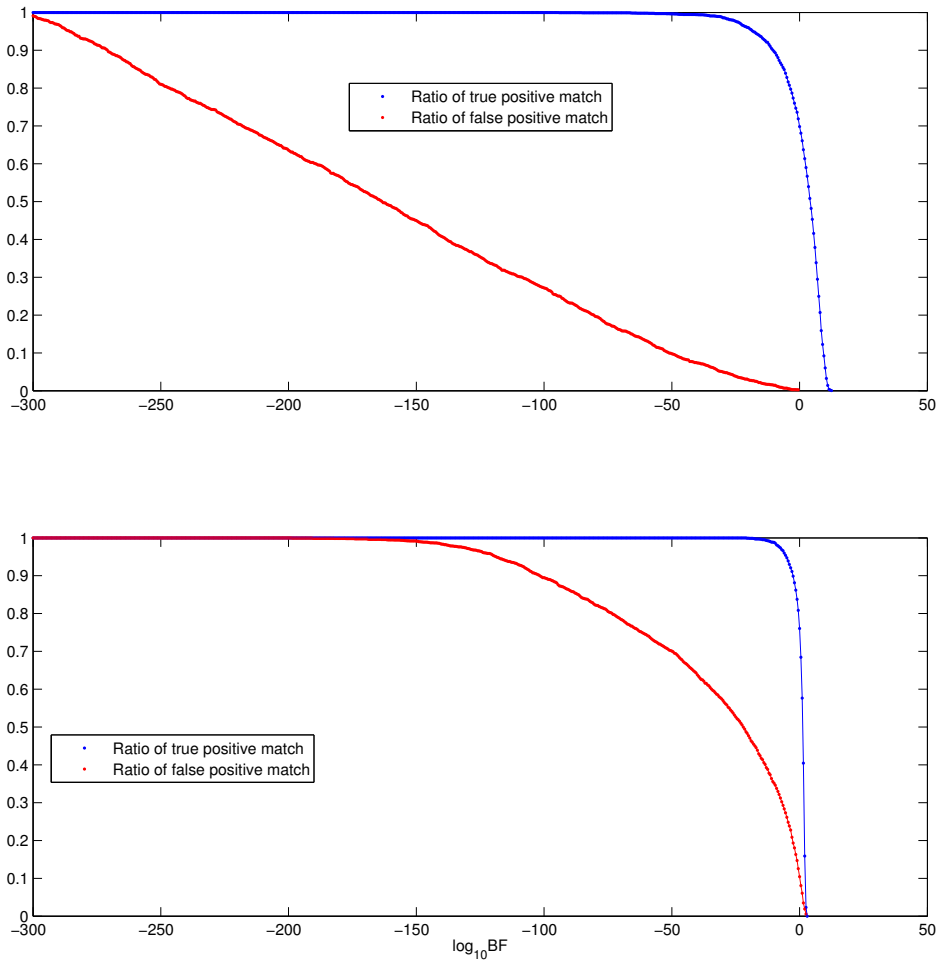

Figure 5.6: Ratio of matches and no-matches for a 2-way (figure above) and for a 10-way (figure below) cross-matching. Bayesian inference approach with Uniform prior. Red line represents the ratio of no matches for the different Bayes factor thresholds and blue line corresponds to ratio of matches. For a positive realistic value of the $log_{10}(B_{ph})$ threshold of 10 the ratio of matches for a real COSMOS matched catalogue is very high whereas the ratio of no matches is very low. This result demonstrates a good capability of this approach to classify photometric matches (from [Marquez et al., 2014]).

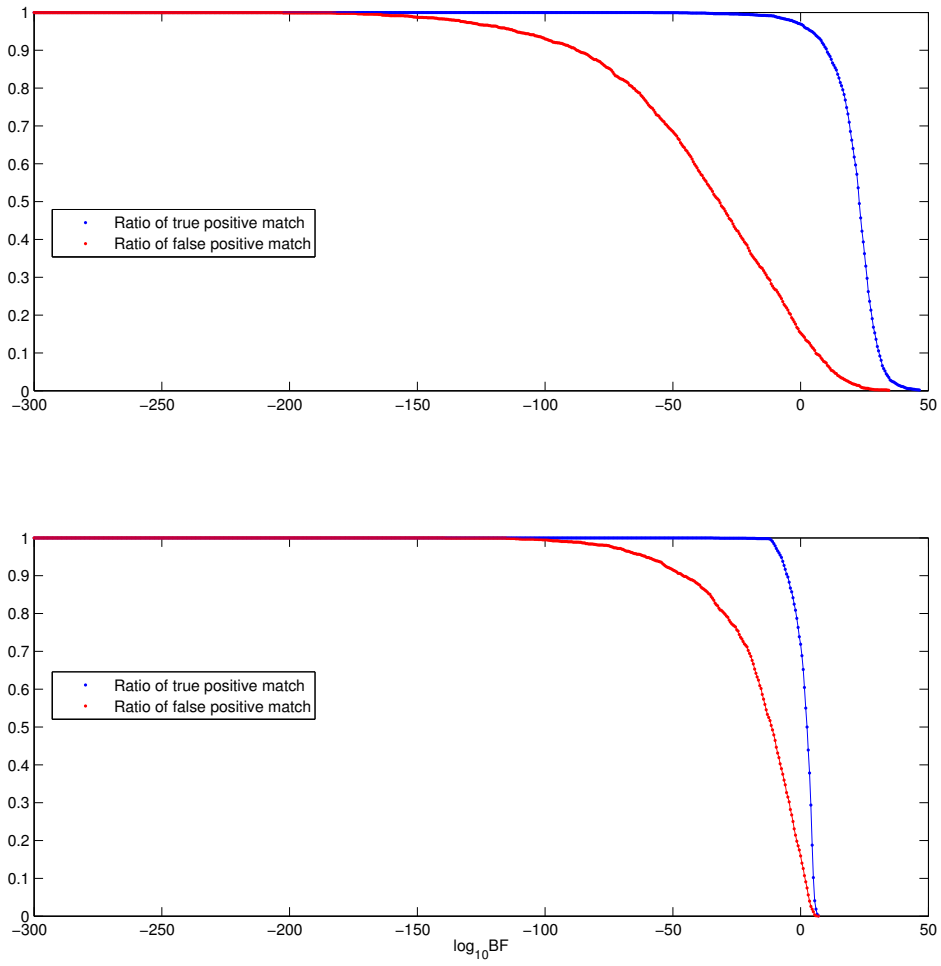

Figure 5.7: Ratio of matches and no-matches for a 2-way (figure above) and for a 10-way (figure below) cross-matching. Bayesian inference approach with flux prior. Red line represents the ratio of no matches for the different Bayes factor thresholds and blue line corresponds to ratio of matches. For a positive realistic value of the $log_{10}(B_{ph})$ threshold of 10, the ratio of matches for a real COSMOS matched catalogue is very high, whereas the ratio of no matches is moderately low. This result demonstrates a moderately good capability of this approach to classify photometric matches (from [Marquez et al., 2014]).

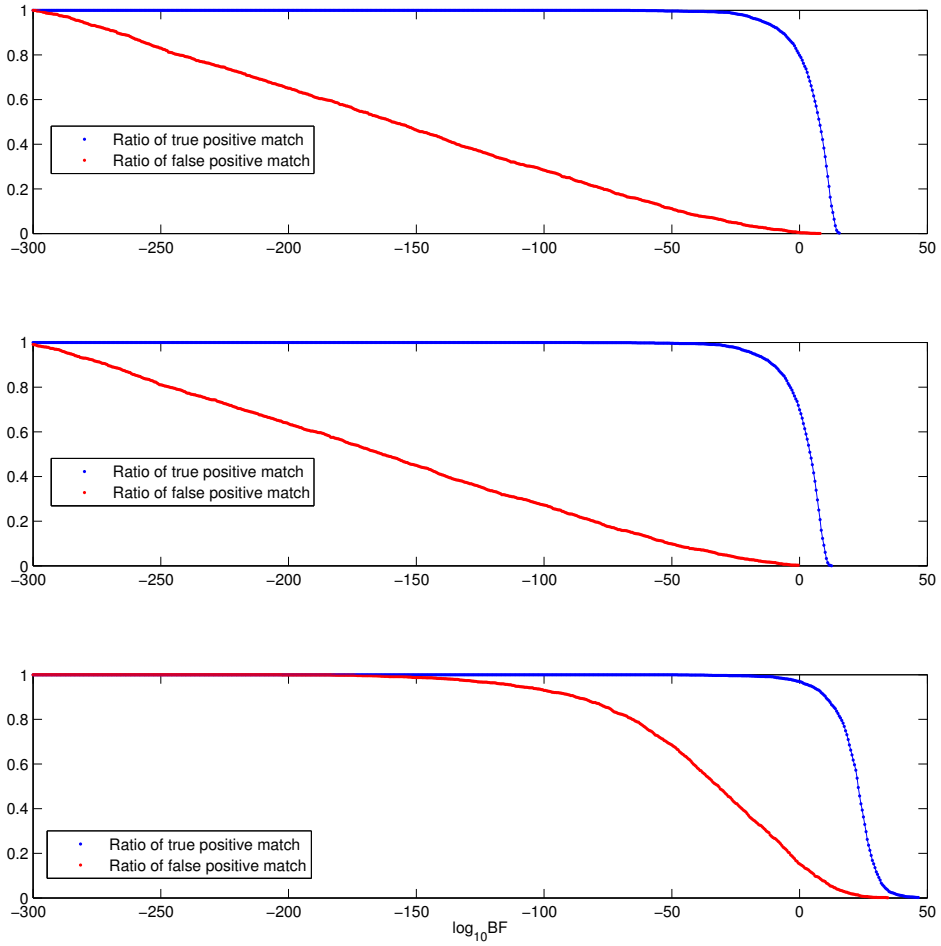

Figure 5.8: The three figures are the result of using a 10 way cross-matching approach for the ten bands of the COSMOS photometric catalogue referred in this chapter. The figure at the top represents the ratio of true positive matches (in blue) and false positive matches (in red) when using the surface brightness prior, as detailed in Section 3.6.4. This result is very similar to the one of the figure in the middle which corresponds to the ratio of true and false positive matches (same colour code) when using a Uniform prior. The figure at the bottom shows the corresponding results when using flux prior. In this last case, for a value of $\log_{10} BF$ of larger than 2, the ratio of false positive matches is higher than in the two other cases (from [Marquez et al., 2014]).

assessment of the method. In general, in terms of capability of matching discriminators, the results from the 10-way approach show the superior quality of the Bayes factor, better than the result from the 2-way approach. This proves that the Bayesian inference allows for a progressive improvement based on ingesting additional information into the framework.

The comparison of the ratio of true and false matches against the $\log_{10} BF$ is shown in Figure

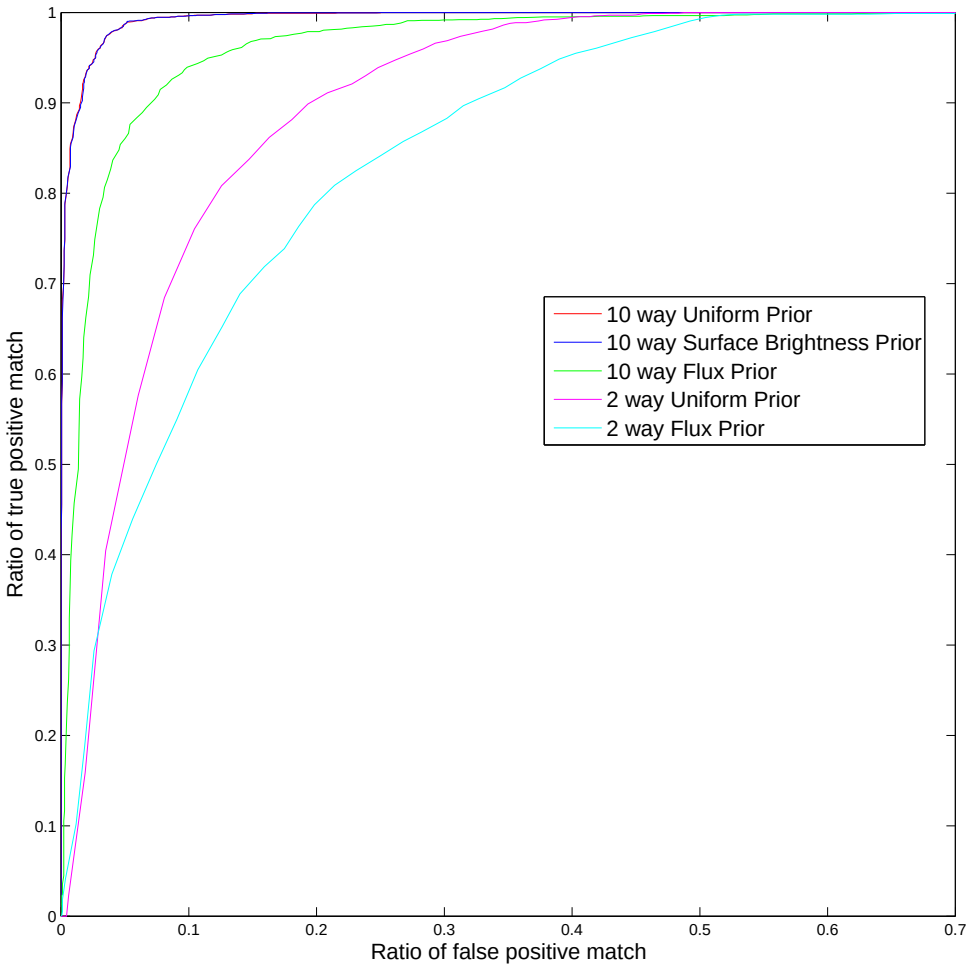

Figure 5.9: The Receiver Operating Characteristic (ROC) curve for the use of different priors in the photometric matching problem. This curve shows the true positive match ratio versus false positive match ratio.
The area under the curve varies with the different priors; the larger the area, the better the performance: Uniform prior (red), flux prior (blue), surface brightness prior (green), 2-way Uniform prior (magenta),2-way flux prior(cyan)(from [Marquez et al., 2014]).

5.8 for the three priors considered in this research - Uniform, flux and surface brightness priors.

We observe from these results that the ratio of false positive matches against true positive matches differs depending on the prior used and on the Smooth Factor applied. From these results we also conclude that a Smooth Factor of 0.03 seems to reach an acceptable and balanced compromise in terms of the model and data uncertainties and their influence on the classification.
Regarding the different priors tested and considering the 2-way approach as an implicit prior, we can see in figure 5.9 that the three 10-way approach produce results of better quality for the three priors than the equivalent 2-way approach. In effect, the area under the ROC curve is considerably smaller for the 2-way approach than for the 10-way, showing that the 2-way approah yields more cases of false positive match than the 10-way approach.; then within the 10-way approach, the one with a Uniform prior seems to achieve the best quality, followed by the surface brightness prior and then the flux prior. It seems that the reduction in computation brought about by the flux prior is penalizing the quality of the discrimination between true positive cross-matches and false positive cross-matches.
Two main observations can be derived from these results: the influence of the different sources of uncertainty plays a key role in the outcome of any valid method of extracting information from measured data. Therefore a detailed consideration of the uncertainties of the data and of the model is crucial in order to obtain realistic outcomes when using a Bayesian inference framework in data analysis. Another observation is the fact that different priors yield different results but the overall balance of the Bayes factor between quality of the fit and complexity of the model is retained. It can be observed that the approach based on the flux prior produces results of worse quality compared to Uniform priors and priors based on surface brightness, in terms of the Bayes factor classifying the cross-matched versus non cross-matched galaxies. However, when combined with the astrometric Bayes Factor, these differences are expected to be mitigated, alleviating one of the disadvantages of the photometric Bayes factor in terms of its sensitivity to the prior.
In terms of computational cost, the solution with a 3D Uniform prior (either l-way or n-way approach) is more demanding due to the fact that the number of integrals to be computed is higher compared to the Flux prior solution.
The photometric cross matching solution presented here yields a characterization of the galaxies through the SED parameters fitting, i.e. redshift and template type. Figures 5.10 to 5.15 show the distribution of redshift and template type for the COSMOS Tile78 with smooth factor 3 and the three main priors - Uniform, Surface Brightness and Flux. Figures 5.16 and 5.17 show the distribution of redhift and SED templates respectively from the COSMOS zphot catalogue. We observe a clear similarity in the distributions of both redshifts and SED templates for the Uniform and Surface Brightness priors, as expected. However, the results with the Flux prior are largely different. This result is in good alignment with the validation outcomes presented in Figures 5.6 to 5.8. Based on those results, we can conclude that the Flux prior is less reliable than the solutions with Uniform or Surface Brightness priors. Also, when comparing the redshift and SED templates distribution of the COSMOS zphot catalogue with the results from our GCES phtometric cross matching, we observe that our GCES system allows a SED fitting with a much wider spectrum of SED parameters (the GCES histograms are wider). This can be explained because the SED grid used by our GCES system has undergone a redshift binning towards higher redshifts. This means that the granularity of our SED grid is considerably higher than the one used in the catalogue and therefore our photometric cross matching algorithm is capable of fitting SED galaxies of a wider redshift spectrum.

A total of 6242 tuples with dubious quality (low quality flags) were identified in COSMOS. The photometric crossmatching yielded around 15% of those tuples as true matches and it would achieve a good SED fitting against the COSMOS SED model library. For the rest of the tuples, the photometric Bayes Factors were low, and this result confirms the low quality indicated in the COSMOS catalogue.
Figure 5.18 shows the distribution and frequency of logarithm values of the photometric crossmatching Bayes factor obtained for the COSMOS Tiles with low quality flags in the bands Subaru B, V, I

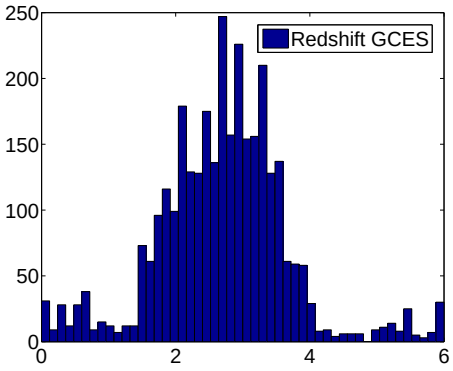

Figure 5.10: COSMOS Tile 78 matched tuples when prior is Uniform and Smooth factor is 3. Distribution of redshift.

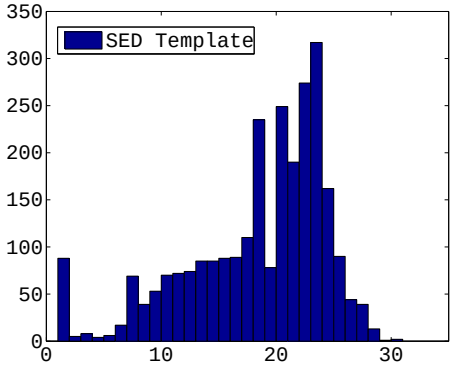

Figure 5.11: COSMOS Tile 78 matched tuples when prior is Uniform and Smooth factor is 3. Distribution of SED templates.

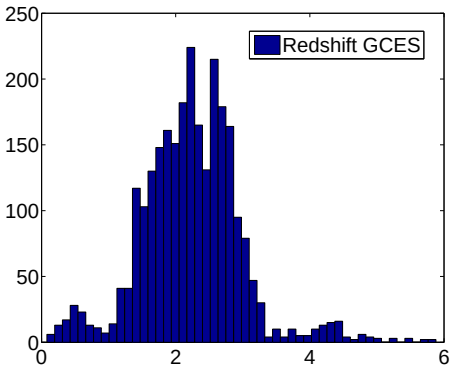

Figure 5.12: COSMOS Tile 78 matched tuples when prior is Surface Brightness and Smooth factor is 3. Distribution of redshift.

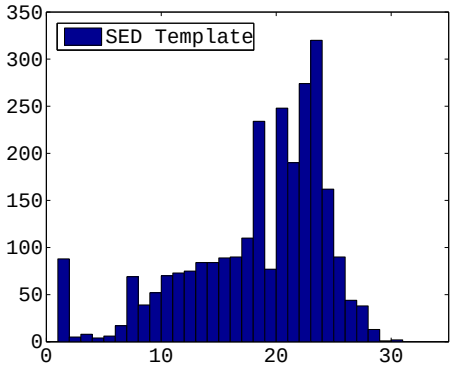

Figure 5.13: COSMOS Tile 78 matched tuples when prior is Surface Brightness and Smooth factor is 3. Distribution of SED templates.

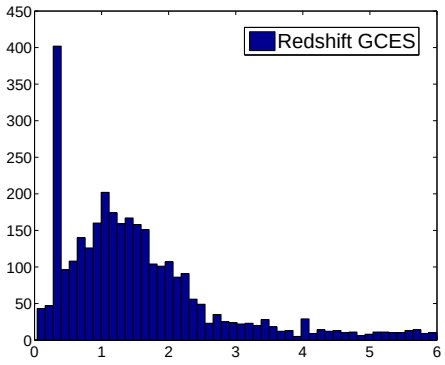

Figure 5.14: COSMOS Tile 78 matched tuples when prior is Flux and Smooth factor is 3. Distribution of redshift.

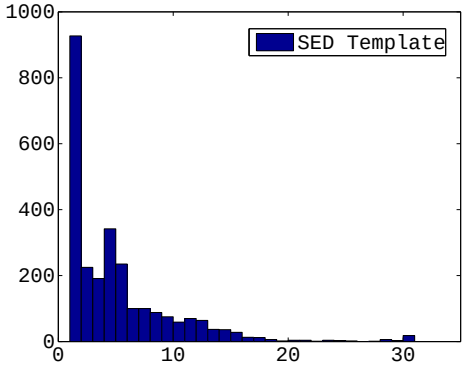

Figure 5.15: COSMOS Tile 78 matched tuples when prior is Flux and Smooth factor is 3. Distribution of SED templates.

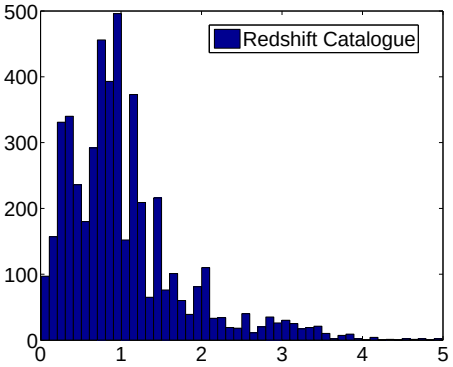

Figure 5.16: Distribution of redshifts. COSMOS zphot catalogue. Tile 78.

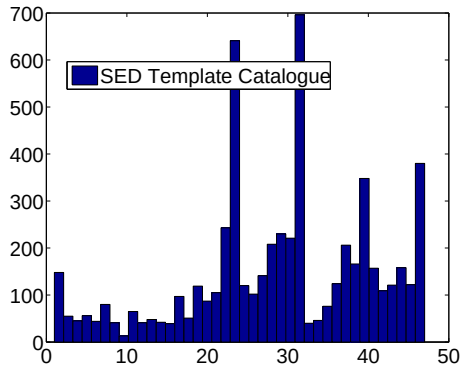

Figure 5.17: Distribution of SED templates. COSMOS zphot catalogue. Tile 78

and Z. This distribution shows, as expected, that the majority of the tuples have a low photometric Bayes factor.

Figure 5.19 shows the cumulative distribution of photometric Bayes Factor values for the dubious

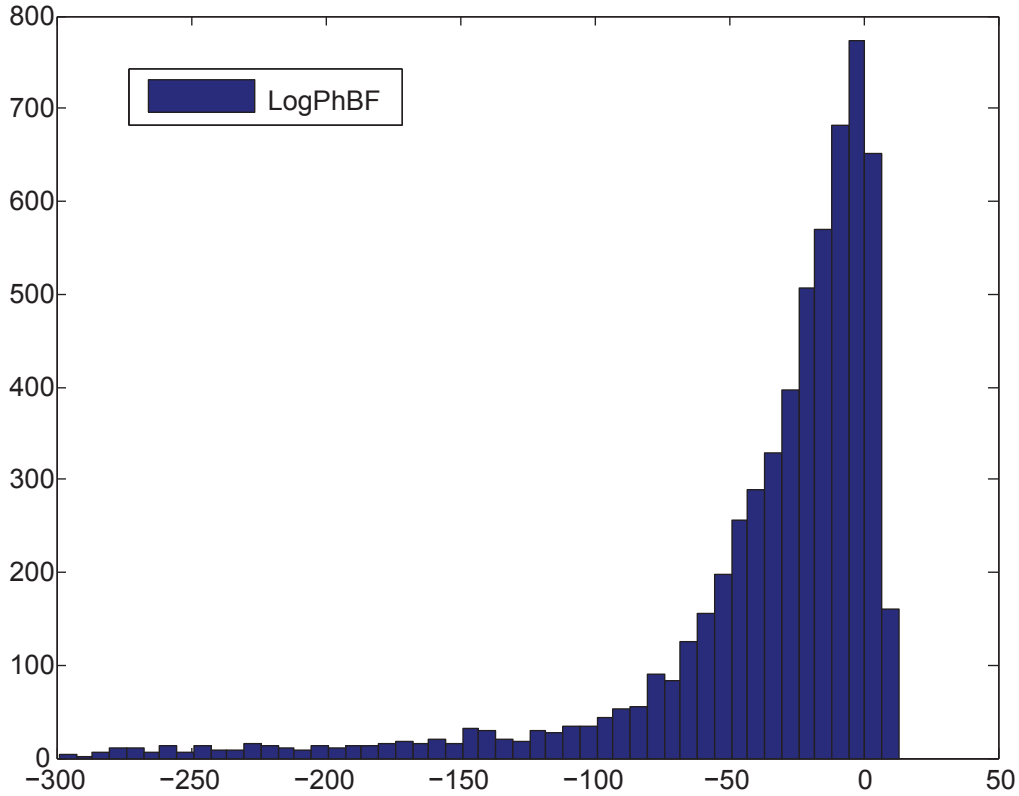

Figure 5.18: Distribution and frequency of the photometric Bayes factor in logarithmic terms for the COSMOS tuples with low quality (dubious cross-matching).

areas of COSMOS.

5.3.2 Source Contour Extraction

It is important to note that due to the dimension of COSMOS catalogue, we extracted for this research a statistical representative subset of COSMOS data. Using the Matlab random function (validated in Chapter 3) we selected randomly 1000 sources from COSMOS Tile 78. Selecting this 1000 sources in each of the 10 bands gives a total of 10000 contours which went through the surface

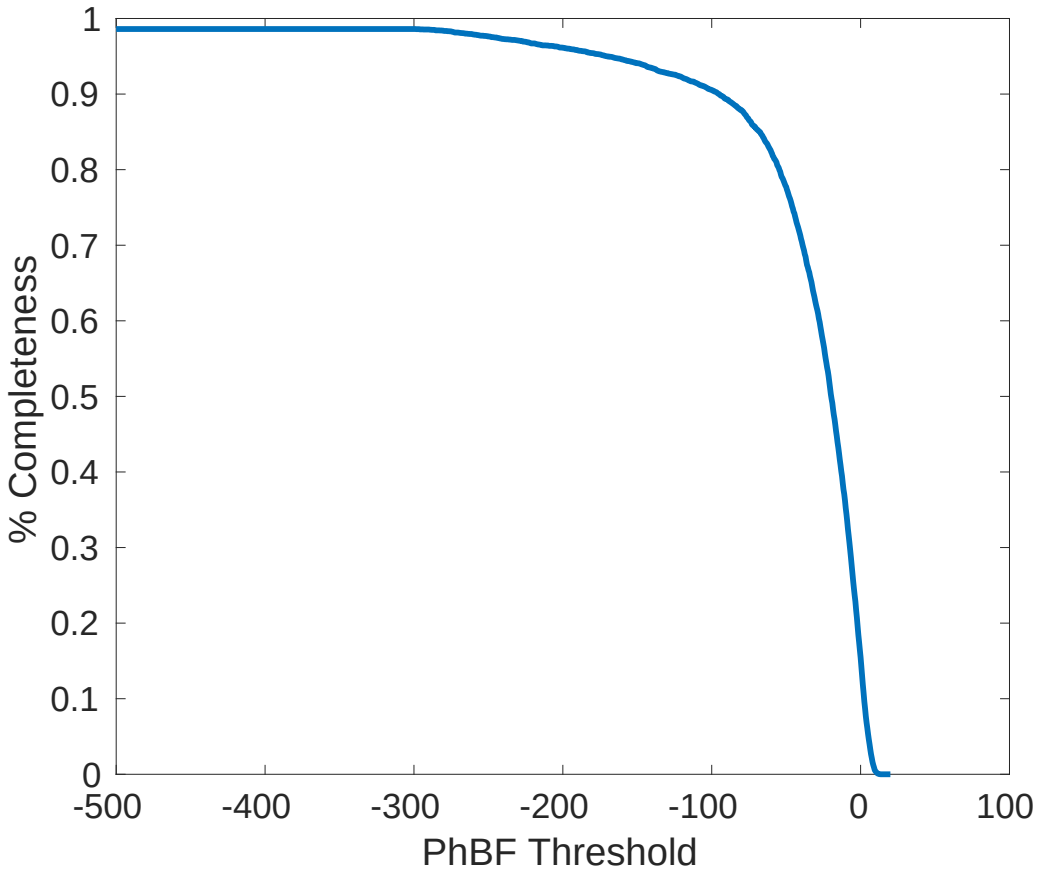

Figure 5.19: Cumulative Distribution Function of PhBF values for dubious cross-matching.

brightness computation of the expert system. This process yielded active contours, counts and the associated AB magnitudes. With these values, and considering the general results characterization presented in both the COSMOS catalogue release of 2008 and more recently in 2016, we performed a similar applicable characterization with the values from the expert system.

Depth in AB Magnitudes In Figure 5.20 the maximum values of AB magnitudes are represented. For this study, in the figure, we have not considered those sources from COSMOS catalogues and with a value of AB Magnitude above 40 because they are not representing the majority of the population and these values are possibly coming from artifacts or some problems in deriving the COSMOS catalogue AB magnitudes. These values for AB magnitudes were not found from the expert system results. The Figure represent those maximum values in each band from the COSMOS catalogue and the expert system (GCES).
The X axis of the graphic is ordered in terms of wavelength, therefore the order of bands represented are: CFHT U, SUBARU B, g, V, r, i, CFHT i, SUBARU z, UKIRT J and CFHT Ks.

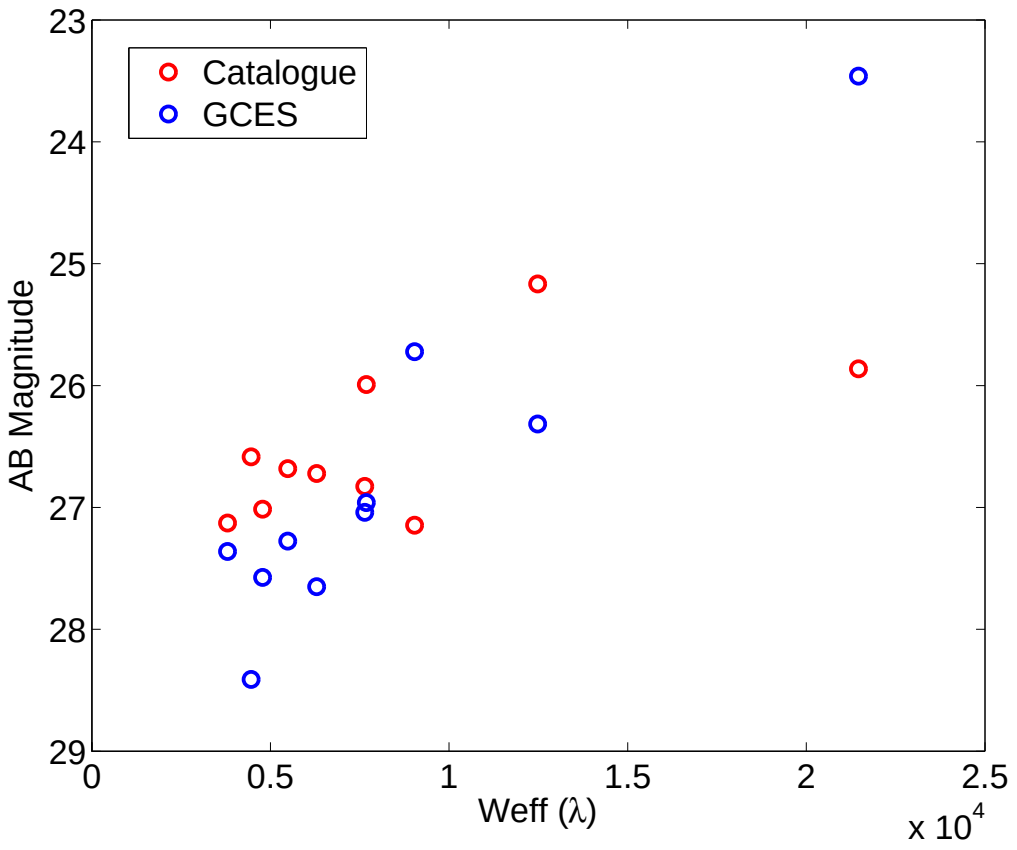

Figure 5.20: Limiting AB Magnitude for COSMOS catalogue bands:CFHT U, SUBARU B, g, V, r, i, CFHT i, SUBARU z, UKIRT J and CFHT Ks

Completeness Figures 5.21 to 5.40 represent the estimated completeness of each band as a function of the magnitude, for both cases, COSMOS catalogue and GCES. The differences in AB magnitude values between COSMOS catalogue and GCES are also represented. It is important to note there that the values of the catalogue indicating no measurement or problems in the measurement $(99, -99)$ are excluded. This is the reason why the COSMOS catalogue completeness figures show a maximum around 80% of completeness instead of approaching 100%.
It can be observed that not all bands under consideration exhibit the same behaviour in terms of the completeness curve. In effect, the completeness of GCES is better than the one for COSMOS catalogue for the SUBARU B band, where the COSMOS release paper ([Capak et al., 2007]) has declared known calibration problems. This could mean that our expert system GCES is capable of coping with calibration problems without manual intervention. It can also be observed that the difference in AB values between the COSMOS catalogue and the GCES remains in general very near the zero point.
From magnitude values lower than 25, the completeness of GCES is clearly superior than the one from COSMOS catalogue. Between 25 and 28 magnitude values is where it is expected to find the maximum difference in AB magnitudes for COSMOS catalogue and GCES. This can also be observed in the accompanying figures.
Figure 5.31 presents systematic fainter magnitudes with our GCES system, compared to the values of the COSMOS catalogue. The reason may be found in the fact that the COSMOS catalogue has computed the magnitudes for the Subaru image with improved PSF in that band, whereas our GCES computed the magnitudes from an image without the improved PSF. This choice allowed us to stress the GCES with the less favourable case and at the same time to not introduce additional variability parameters from the input data (the COSMOS survey use did not have the best psf image for all the bands).
The values clearly out of the diagonal for which GCES magnitudes are brighter than the ones from the COSMOS catalogue, correspond to a group of blended sources.
Figures 5.41 and 5.42 show two representative cases of large and small difference in AB magnitude values between the COSMOS catalogue and our GCES system.

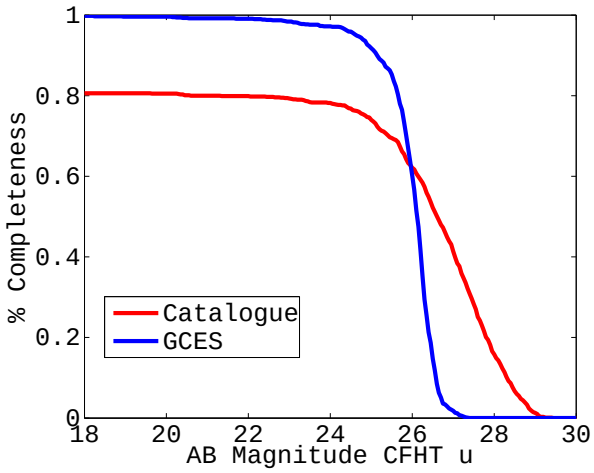

Figure 5.21: CFHT u completeness.

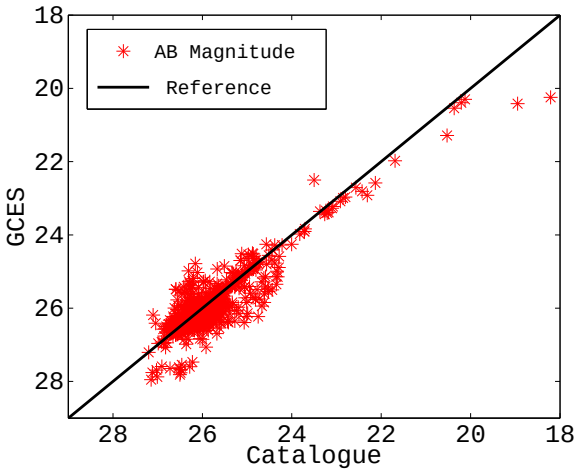

Figure 5.22: Differences between COSMOS catalogue and our GCES system in AB magnitude for CFHT u band.

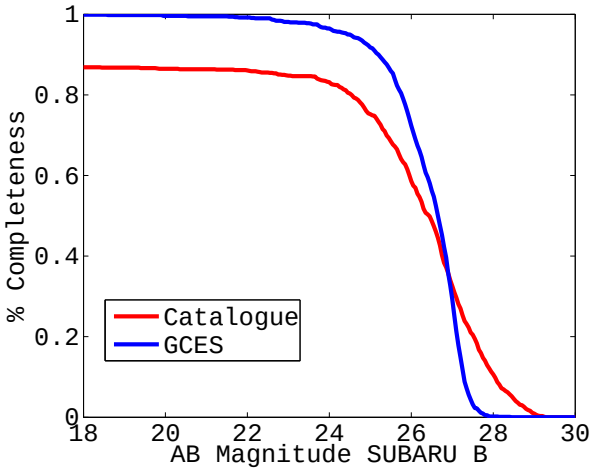

Figure 5.23: SUBARU B completeness.

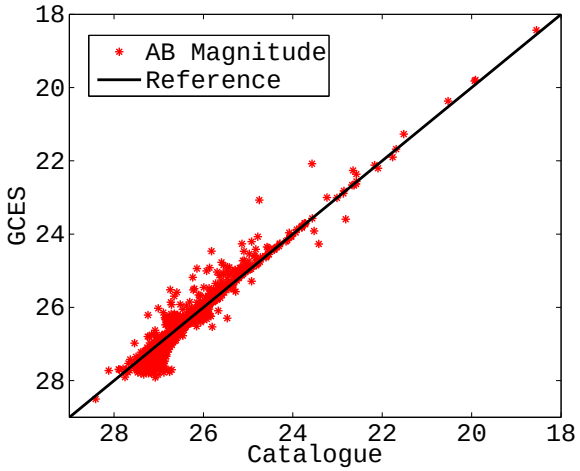

Figure 5.24: Differences in AB magnitude for SUBARU B band.

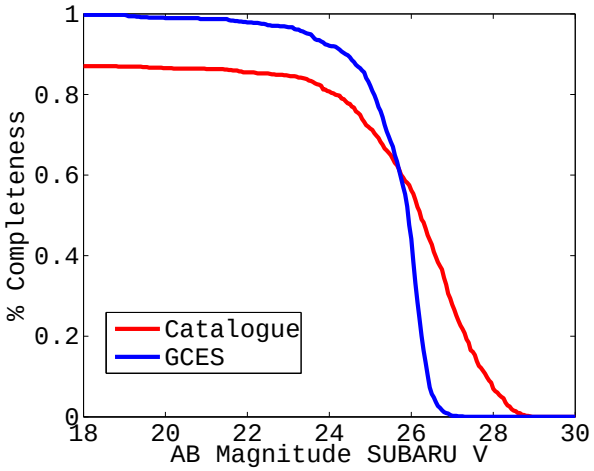

Figure 5.25: SUBARU V completeness.

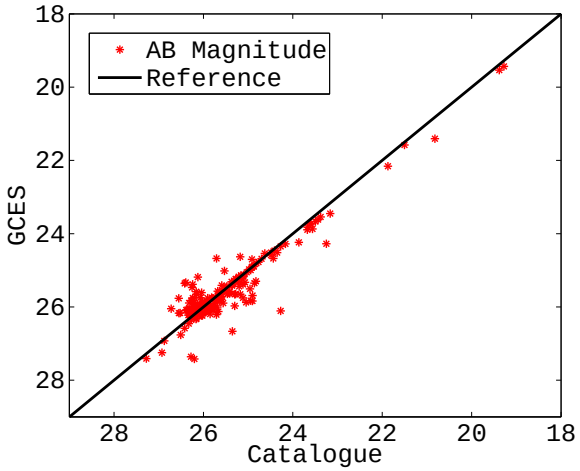

Figure 5.26: Differences in AB magnitude for SUBARU V band.

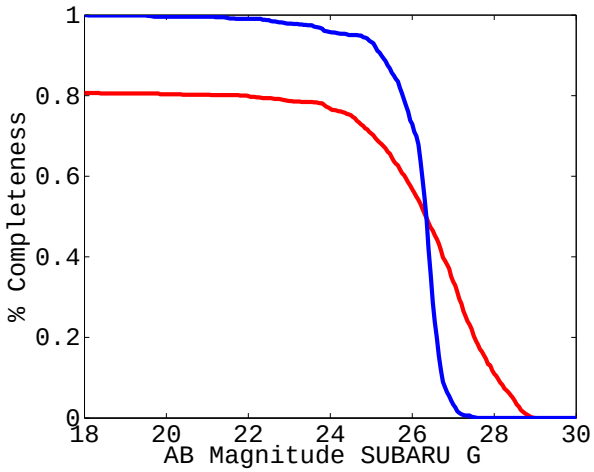

Figure 5.27: SUBARU g completeness.

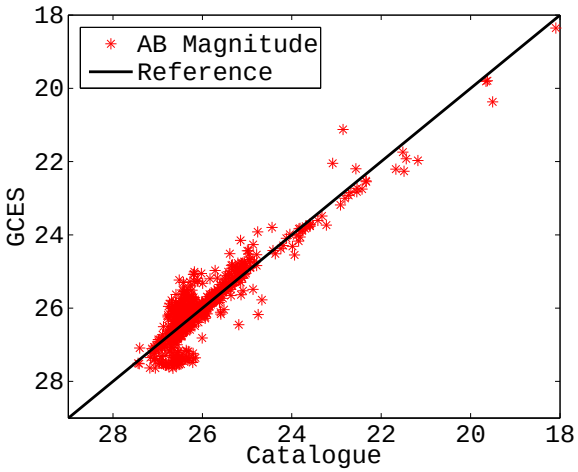

Figure 5.28: Differences in AB magnitude for SUBARU g band.

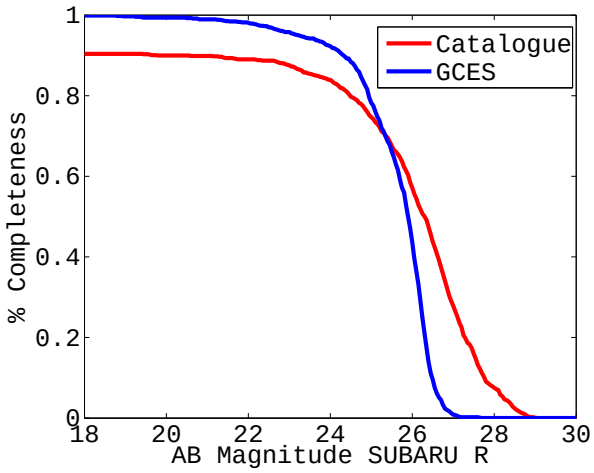

Figure 5.29: SUBARU r completeness.

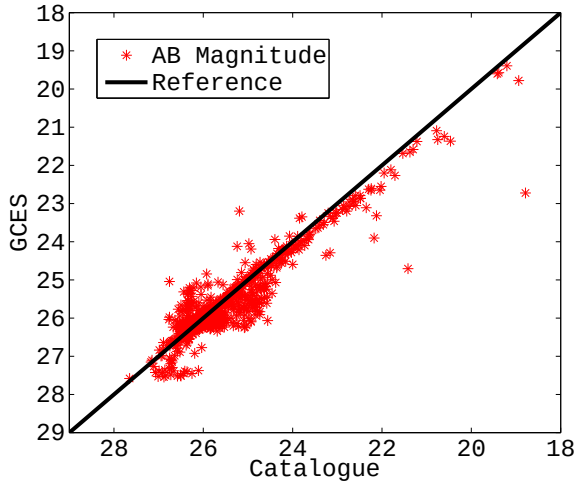

Figure 5.30: Differences in AB magnitude for SUBARU r band.

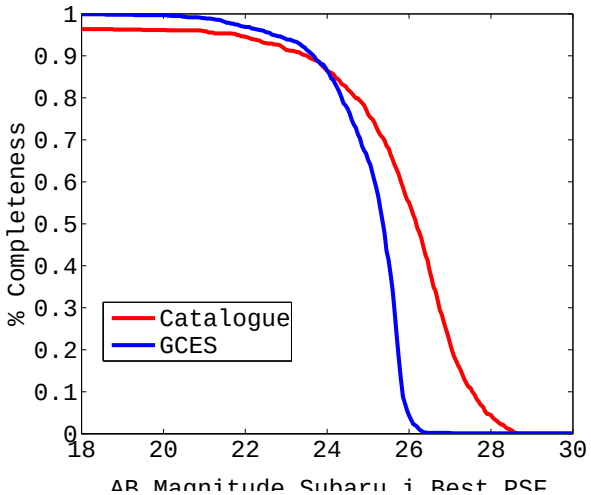

Figure 5.31: SUBARU i completeness.

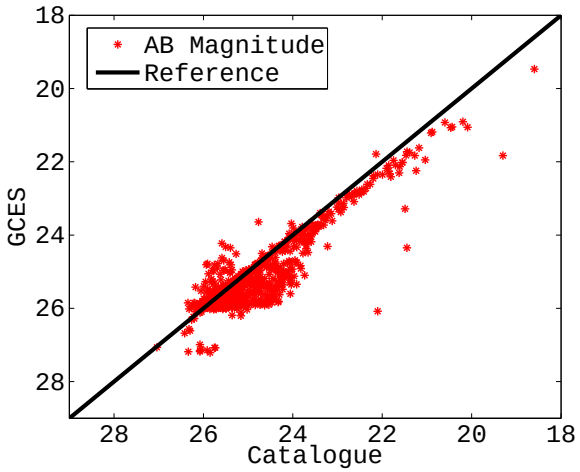

Figure 5.32: Differences in AB magnitude for SUBARU i band.

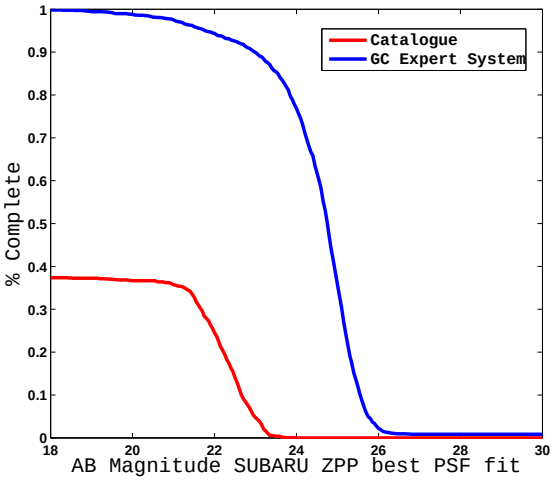

Figure 5.33: SUBARU Z best PSF fit completeness.

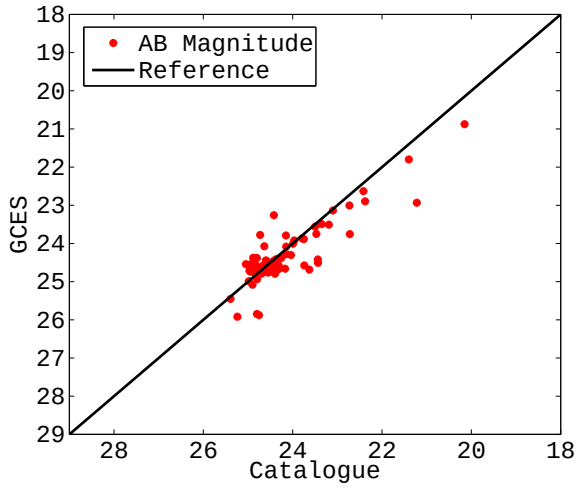

Figure 5.34: Differences in AB magnitude for SUBARU Z band.

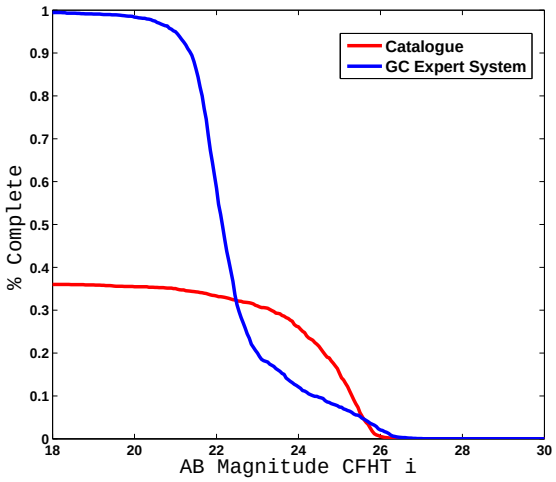

Figure 5.35: CFHT I completeness.

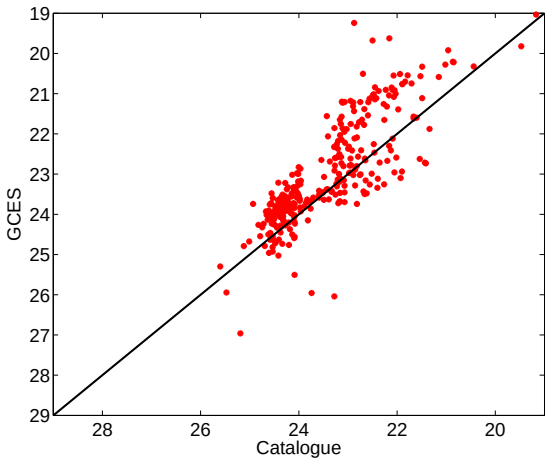

Figure 5.36: Differences in AB magnitude for CFHT I band.

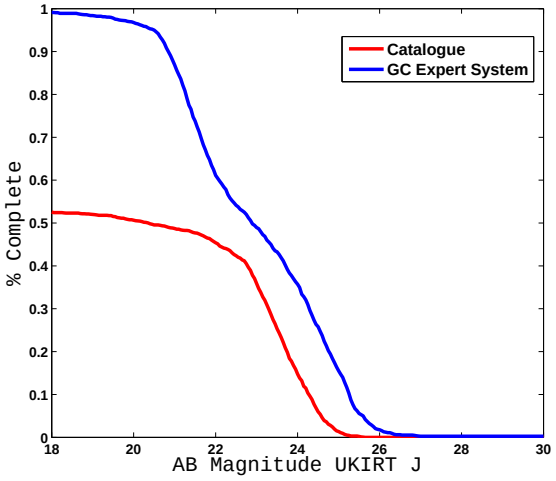

Figure 5.37: UKIRT J completeness.

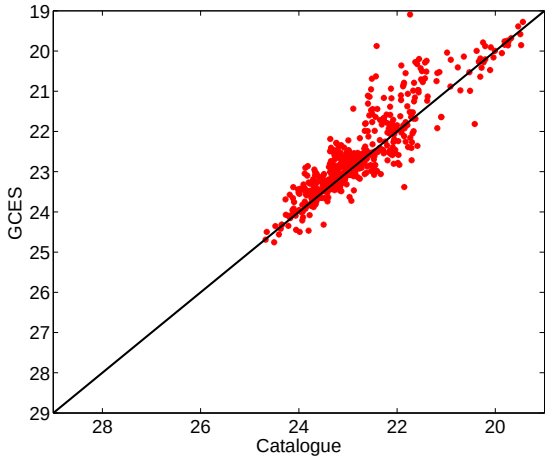

Figure 5.38: Differences in AB magnitude for UKIRT J.

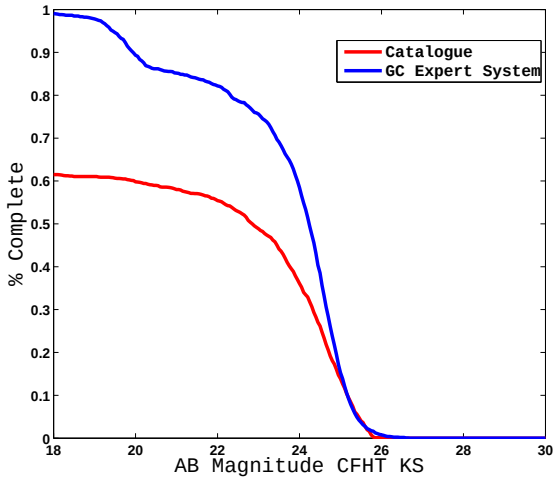

Figure 5.39: CFHT Ks completeness.

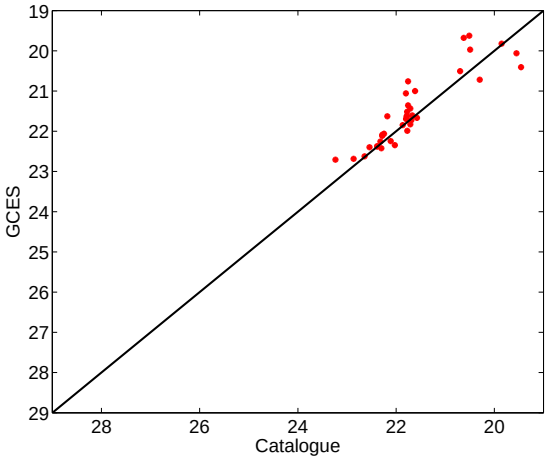

Figure 5.40: Differences in AB magnitude for CFHT Ks.

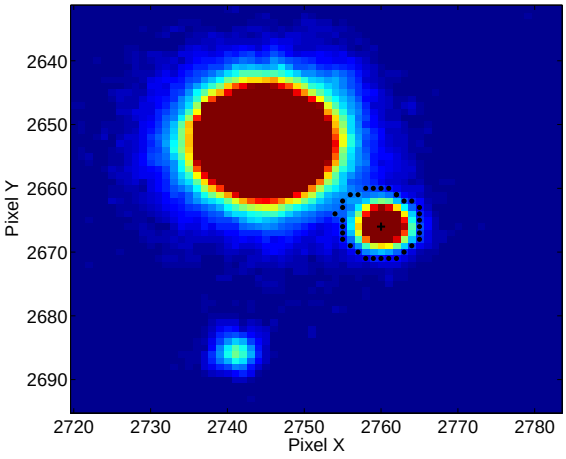

Figure 5.41: AB Magnitude difference of 3.9. SUBARU r.

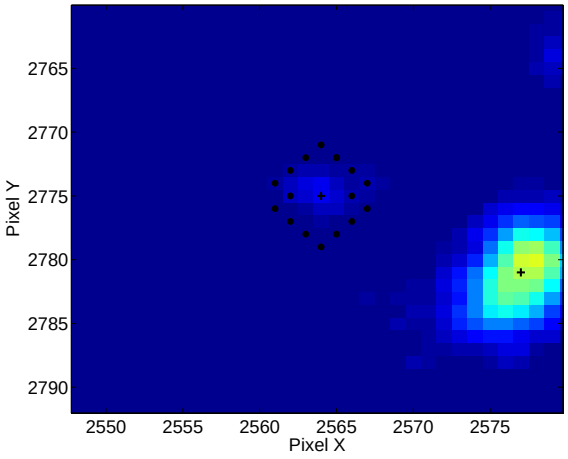

Figure 5.42: AB Magnitude difference of 0.0021. Subaru r.

Color Offset Figures 5.43 to 5.48 represent the scatter and zero-point error for the colour offset between the COSMOS bands – CFHT u, SUBARU B, V, g, r, z, i, CFHT i, UKIRT J, CFHT Ks, and the F814W ACS band for both COSMOS catalogue and GCES AB values. The range of scatter is in good correspondence between COSMOS and GCES, with more dispersion beyond 25 of magnitude in both cases..

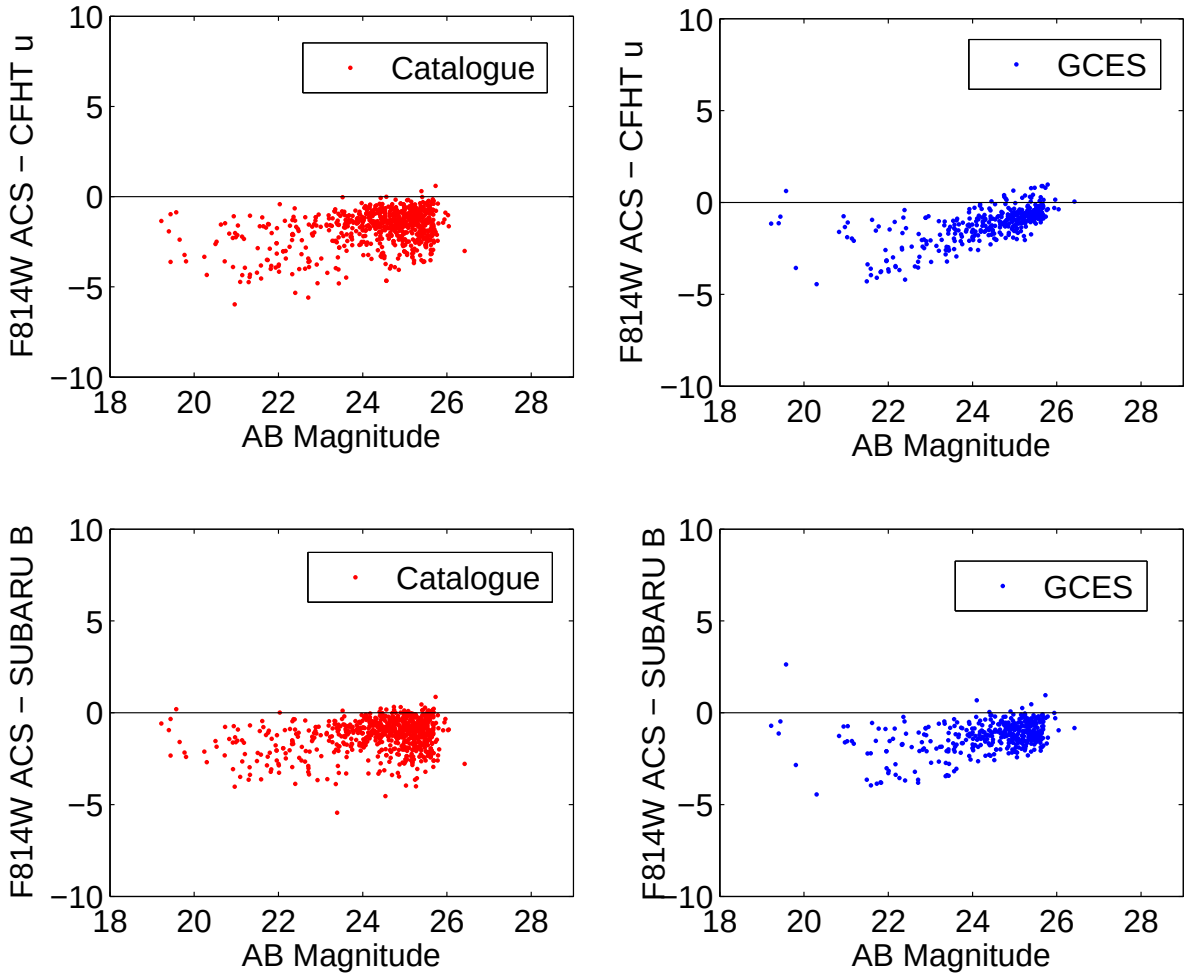

Figure 5.43: Fig a. Colour Offset between F814W and CFHT U from the COSMOS catalogue. Fig b. Colour Offset between F814W and CFHT U from GC Expert system.Fig c. Colour Offset between F814W and SUBARU B from the COSMOS catalogue. Fig d. Colour Offset between F814W and SUBARU B from GC Expert System

Figure 5.48 represents the AB magnitude offset between the COSMOS catalogue and the GC Expert System for the bands SUBARU B, V, r and CFHT i. The scatter and zero-point error for both – the COSMOS catalogue and the GCES system are comparable and in good alignment

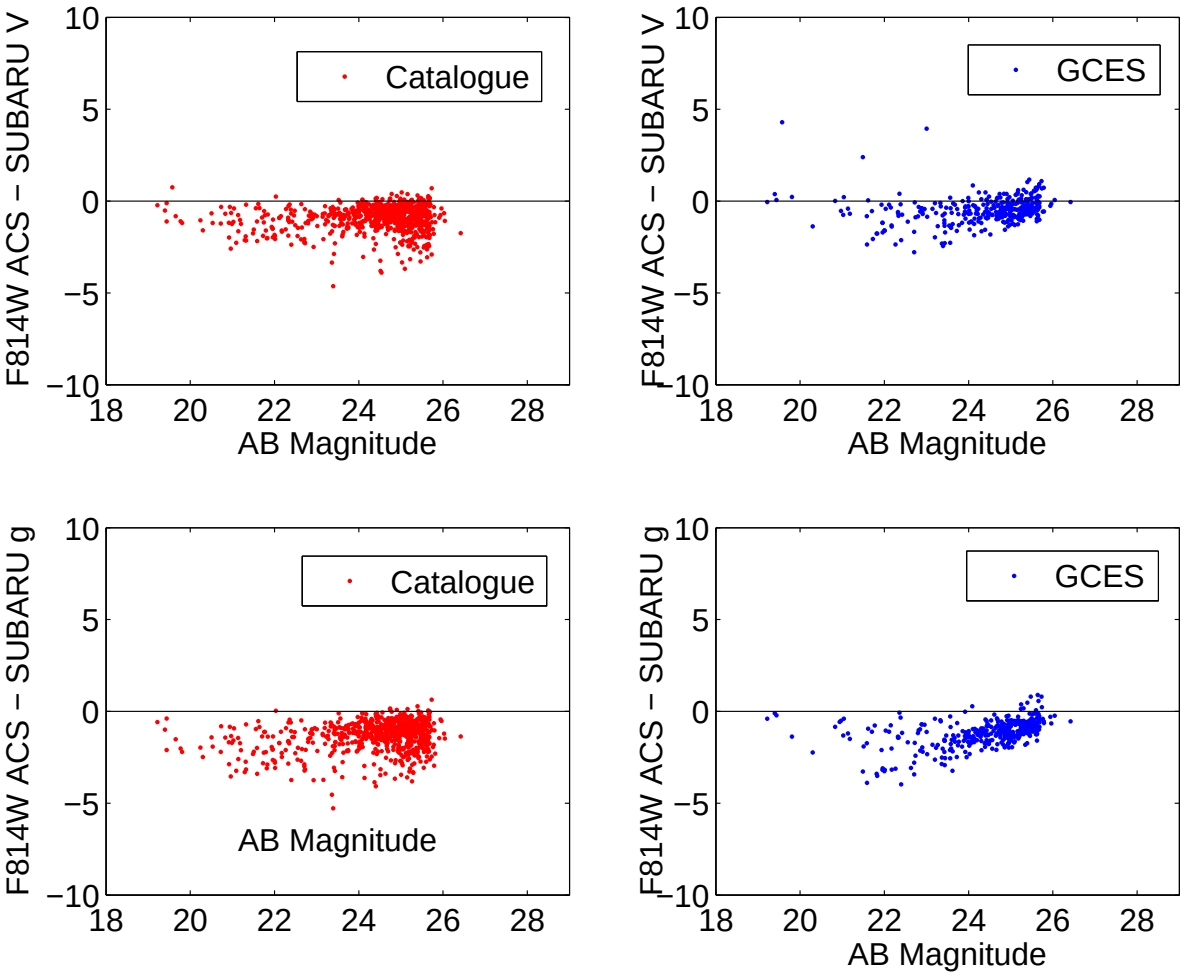

Figure 5.44: Fig a. Colour Offset between F814W and SUBARU V from the COSMOS catalogue. Fig b. Colour Offset between F814W and SUBARU V from GC Expert system.Fig c. Colour Offset between F814W and SUBARU g from the COSMOS catalogue. Fig d. Colour Offset between F814W and SUBARU g from GC Expert System

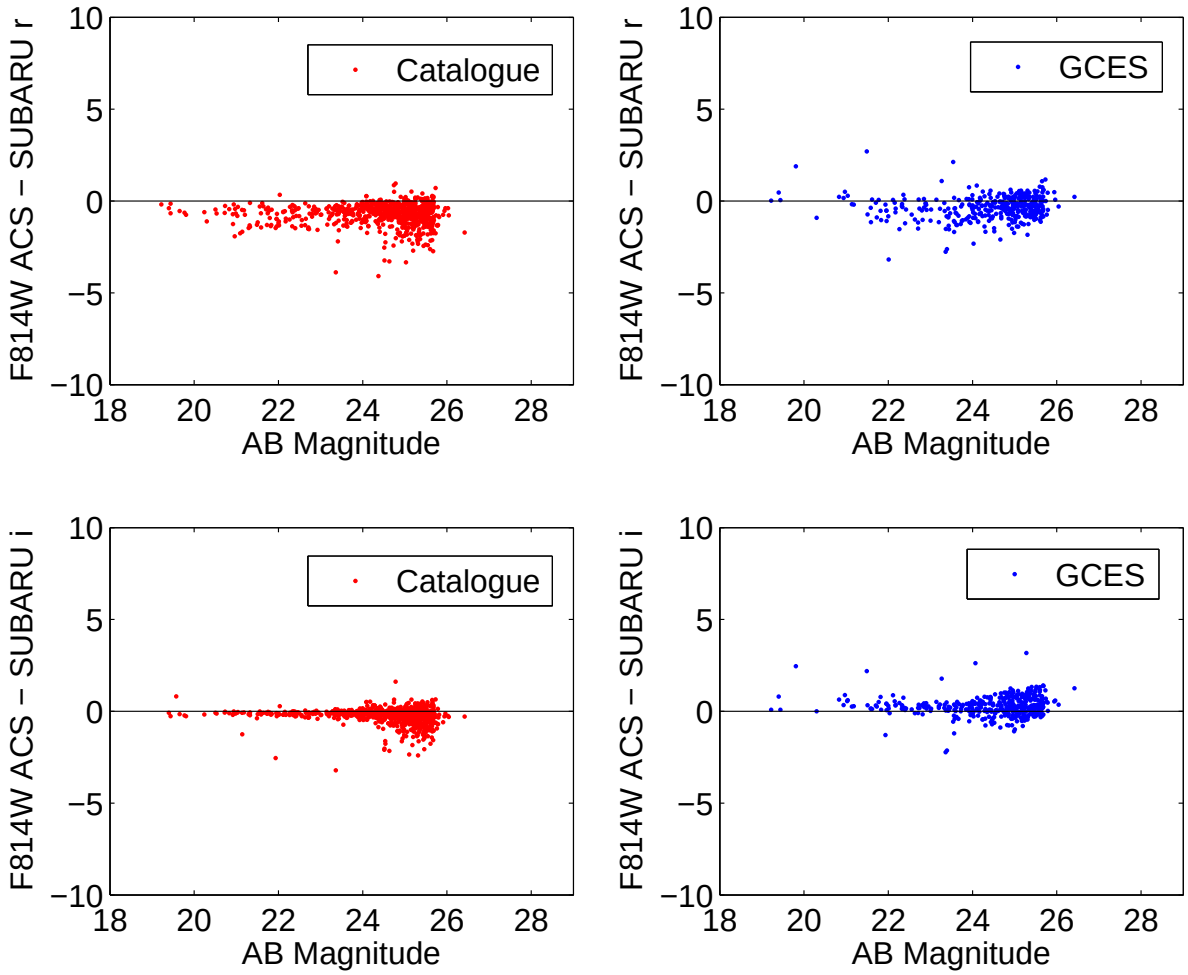

Figure 5.45: Fig a.Colour Offset between F814W and SUBARU R best PSF fit from the COSMOS catalogue. Fig b.Colour Offset between F814W and SUBARU R from GC Expert System Fig c. Colour Offset between F814W and SUBARU I best PSF fit from the COSMOS catalogue. Fig d. Colour Offset between F814W and SUBARU I from GC Expert System

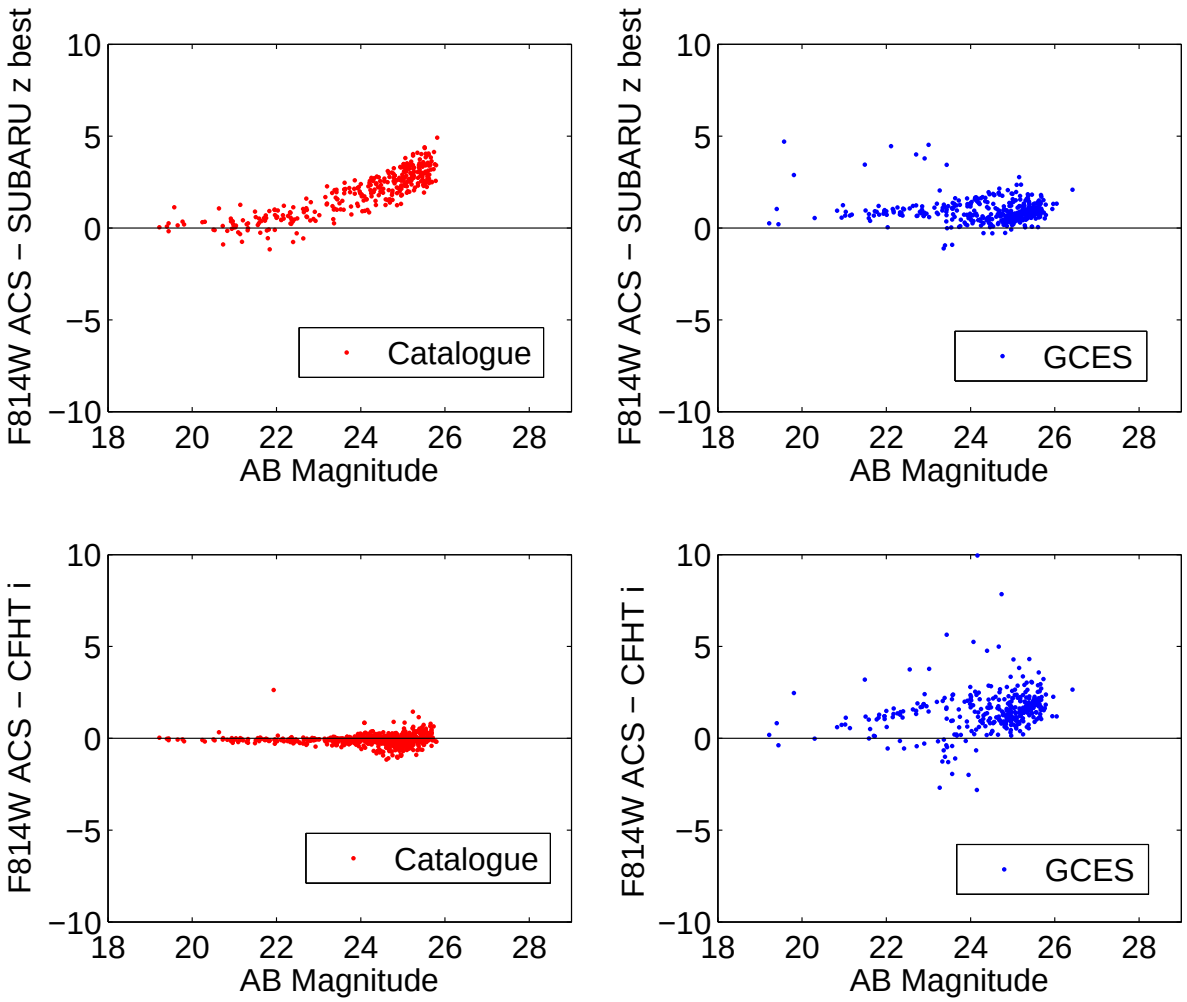

Figure 5.46: Fig a. Colour Offset between F814W and SUBARU z best PSF fit from the COSMOS catalogue . Fig b. Colour Offset between F814W and SUBARU Z best PSF fit from GC Expert System. Fig c. Colour Offset between F814W and CFHT i from the COSMOS catalogue . Fig d.Colour Offset between F814W and CFHT i from GC Expert System.

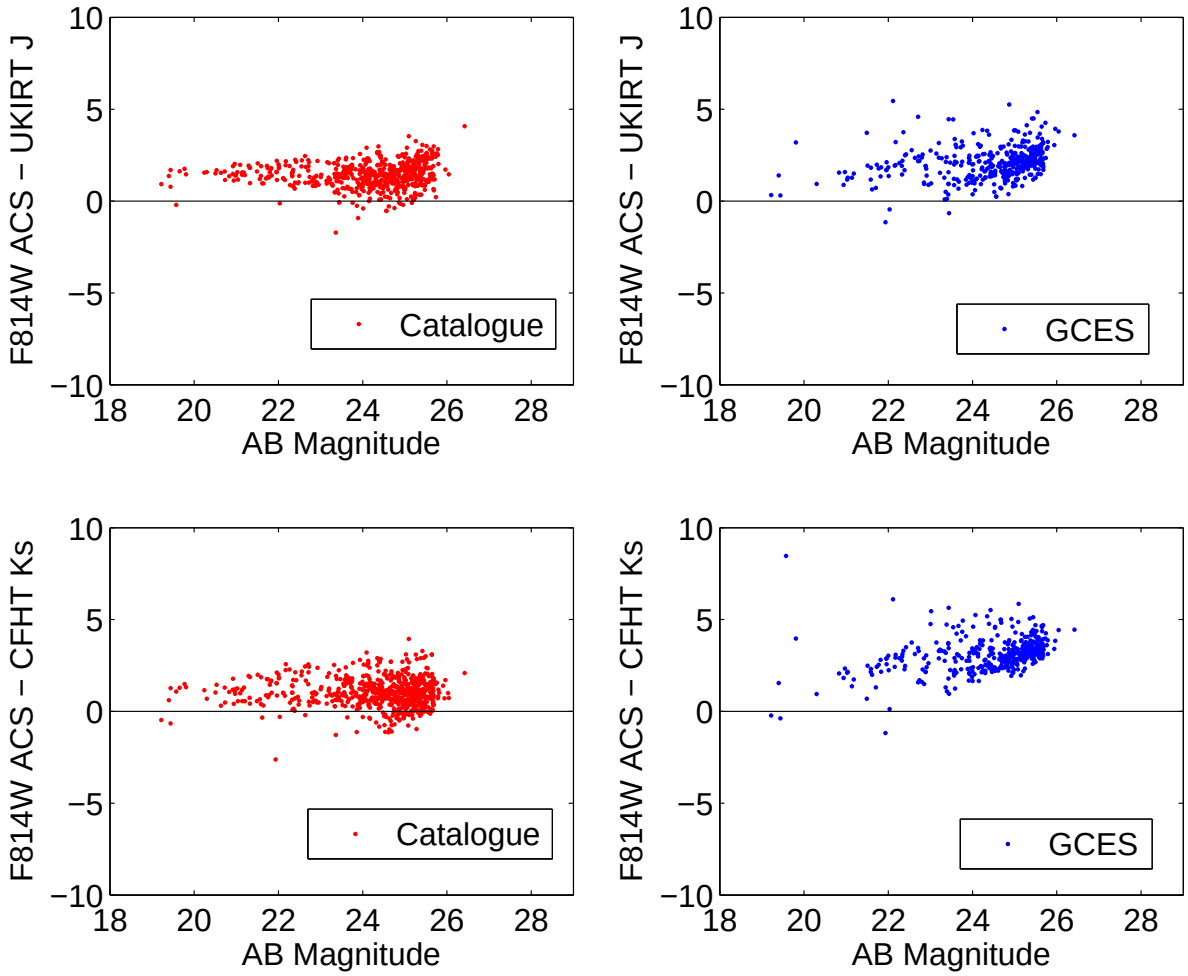

Figure 5.47: Fig a.Colour Offset between F814W and UKIRT J from the COSMOS catalogue . Fig b. Colour Offset between F814W and UKIRT J from GC Expert System . Fig c. Colour Offset between F814W and CFHT KS from the COSMOS catalogue. Fig d. Colour Offset between F814W and CFHT KS from GC Expert System.

with those published in [Capak et al., 2007]. From Figure 5.48 we observe that in all panels, the

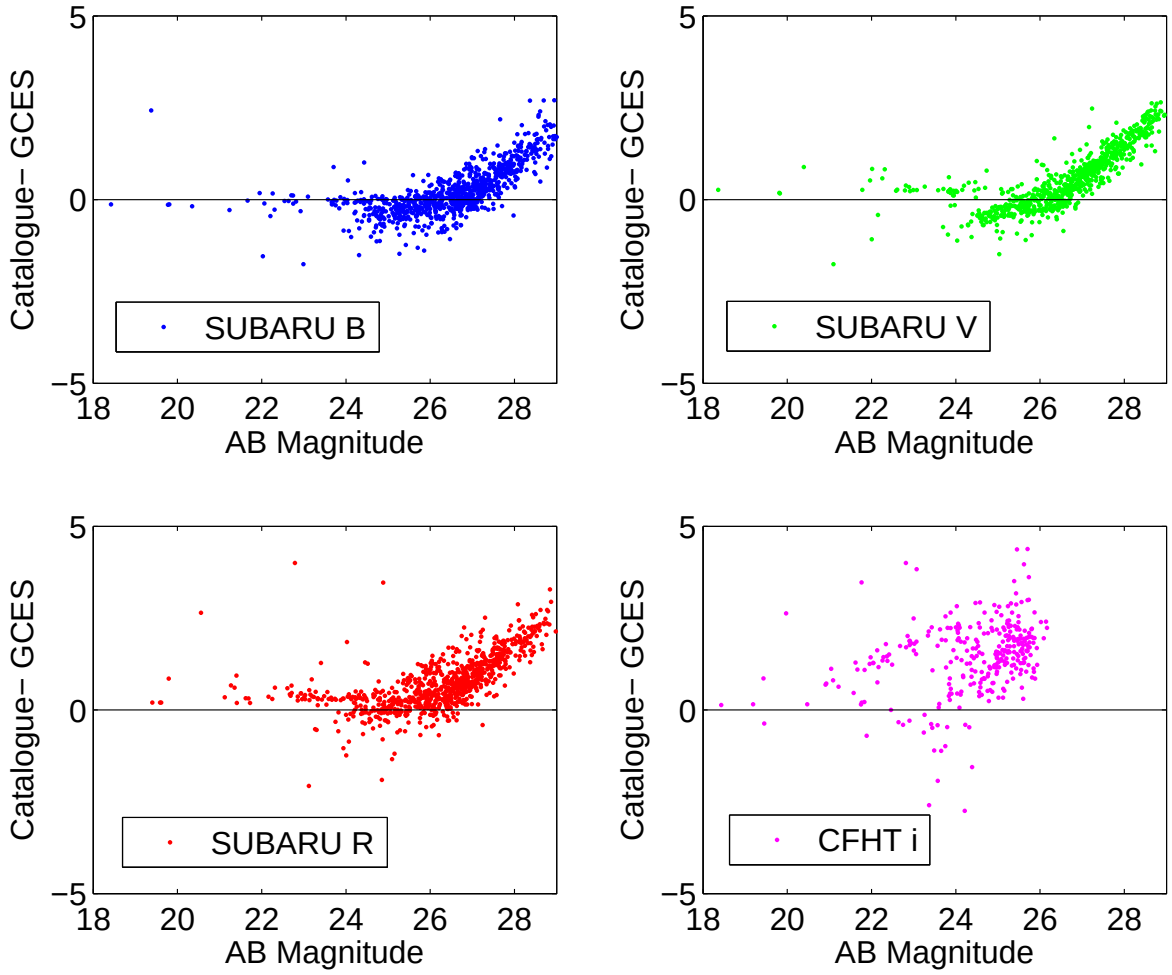

Figure 5.48: Fig a. AB magnitude offset of SUBARU B between catalogue and GC Expert System. Fig b. AB magnitude offset of SUBARU g between catalogue and GC Expert System . Fig c. AB magnitude offset of SUBARU R best PSF fit between catalogue and GC Expert System Fig d.AB magnitude offset of CFHT i between catalogue and GC Expert System .

dispersion from zero increases beyond AB magnitude 25. This is in line with the degeneracy of the photometric values beyond that limit, as reported in [Capak et al., 2007]. Another observation from this figure is that for faint sources, our GCES yields normally brighter values than the catalogue. If we consider the typical scenario of the COSMOS images with crowded areas, it is very common to find faint and bright sources very close. Under this scenario our GCES system would identify a contour beyond the real limits of the faint source and this would result in a false brighter value.

Color-Color Figure 5.49 represents color-color diagrams as the ones of [Capak et al., 2007] but it includes the outcomes from the COSMOS catalogue under study and from GCES. This information will allow us to offer a comparative analysis, as it is described in Section 5.4.2. From the results it is

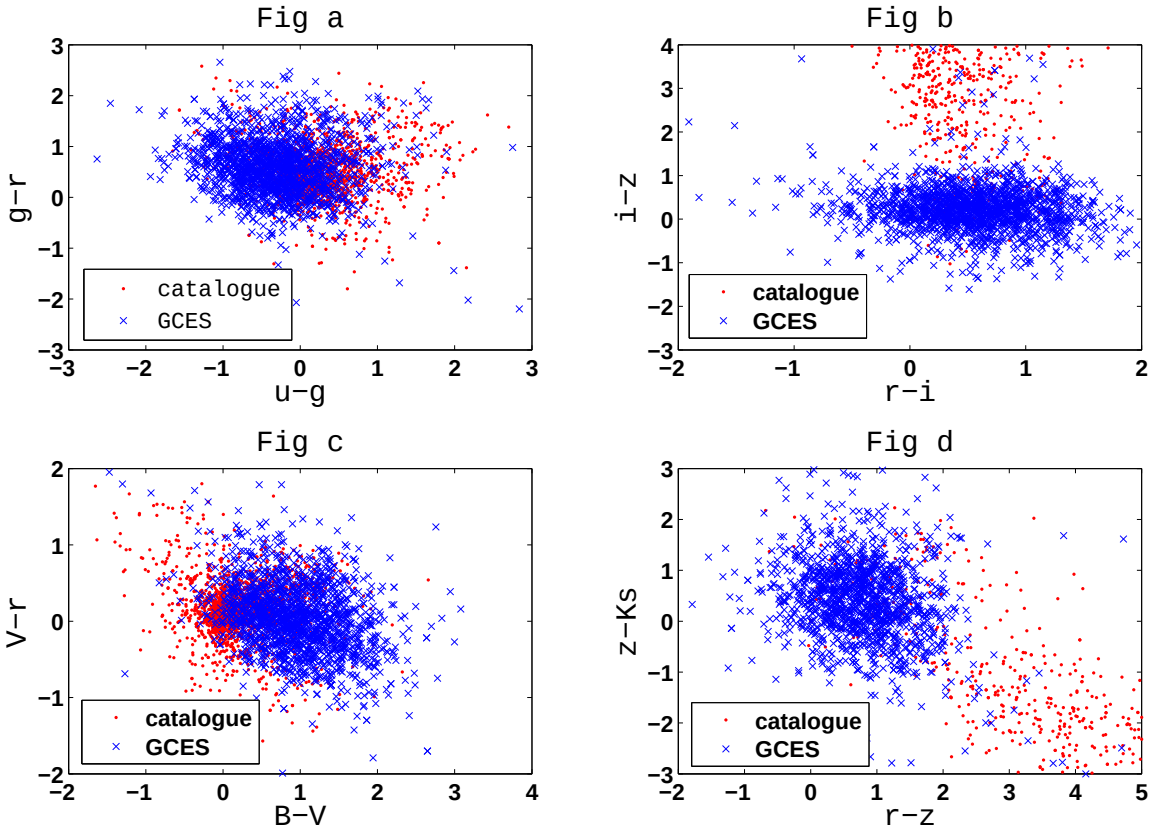

Figure 5.49: Panels a, b, c and d represent various color color diagrams, where the results from the catalogue are shown in red and the results from the GC Expert System (GCES) in blue.

observed that there is more dispersion between COSMOS catalogue and GCES for the cases of i-z and z-Ks than for the cases of g-r and V-r. However and due to the nature of these diagrams, 1000 sources would not seem to be a sufficiently large sample size to allow the drawing of firm conclusions on this.

MATLAB versus Active Contour with real COSMOS Data Figure 5.50 shows the result of the surface brightness computation with the use of MATLAB contouring algorithm on a reduced window of 259 COSMOS real galaxies for the Subaru V band. One important limitation in the assessment of this result, as well as for the results included in this section is the fact that the COSMOS catalogues used do not contain the necessary information to rebuild the contours used to obtain their apparent magnitudes. This is an important limitation for considering the COSMOS catalogue as the reference against which we compare the results from our system. In addition to that, and based on the COSMOS release catalogue publications, it seems that a considerable number of important manual steps are performed before obtaining the final apparent magnitude released in those catalogues; thus, it was not possible to reproduce the steps yielding the apparent magnitudes of the COSMOS catalogues used. From Figure 5.50 we observe that the MATLAB contours leave pixels of the object outside of the contour, leading to systematic fainter sources than the corresponding values from the COSMOS catalogue.

5.3.3 Source Labelling

Figures 5.51 to 5.71 show the Source Labelling characterization in all COSMOS bands under study. The GCES system uses the Voronoi tessellation and a rule based system for the determination of the labelling of the sources (contaminated, partially contaminated or isolated).
Figure 5.51 shows the COSMOS Tile 78 image in the Subaru B band. We have selected a window of 154 sources, marked in white colour, where we ran the labelling process for the ten COSMOS bands under study. This number of sources is statistically representative to demonstrate the capabilities of this labelling process. This restricted window allows us to illustrate a zoom of the whole tile.
In order to ease the visual evaluation of the source labelling system, the images show the contour and the Voronoi tessellation. Recalling the rule based system, in practical terms, the following colour code has been used to distinguish the three cases of labelling: the contour in question is coloured in red if it is contaminating at least one of the neighbour voronoi cells; it is coloured in yellow if at least one contour of a neighbour voronoi cell is contaminating it and, finally, it is coloured in green if its contour is fully contained in its voronoi cell and none of the surrounding contours are contaminating it.
Figure 5.52 shows the result from labelling the 154 CFHT u sources; the active contours were coloured in line with our rule based system: green for isolated sources, yellow for partially contaminated and red for contaminating sources. This result is a preliminary labelling for this band. Similarly, we obtain the source labelling for the rest of the bands, as it can be seen in Figure 5.54 for Subaru B, Figure 5.56 for Subaru V, Figure 5.58 for Subaru g, Figure 5.60 for Subaru r, Figure 5.62 for Subaru i, Figure 5.64 for Subaru z, Figure 5.66 for CHFT i, Figure 5.68 for UKIRT J and Figure 5.70 for CHFT Ks. In all of them, as expected, the source labelling system yields deterministic results. One note of importance here is the fact that the computational accuracy of MATLAB when applying the rule based system may be compromised for those Voronoi cells with at least one vertex in infinity. In order to cope with this apparent limitation of MATLAB, we have descoped from our source labelling syste those cases, which are marked with black stars in the figures.
Figure 5.53 shows the Voronoi tessellation for the CFHT u sources of the COSMOS catalogue for which there is a valid magnitude (different from 99 or -99). The GCES active contour for those sources are represented in white. In black we have included the GCES active contours for those sources for which the COSMOS catalogue has no valid magnitude value. We observe that our GCES can identify a contour on cases for which the COSMOS catalogue cannot. As we progress through the bands the respective figures show black contours where in the previous band the contour was white. This means that the number of source detected varies with the bands, as reflected in the COSMOS catalogue, but our GCES system does not go through the same steps of consideration

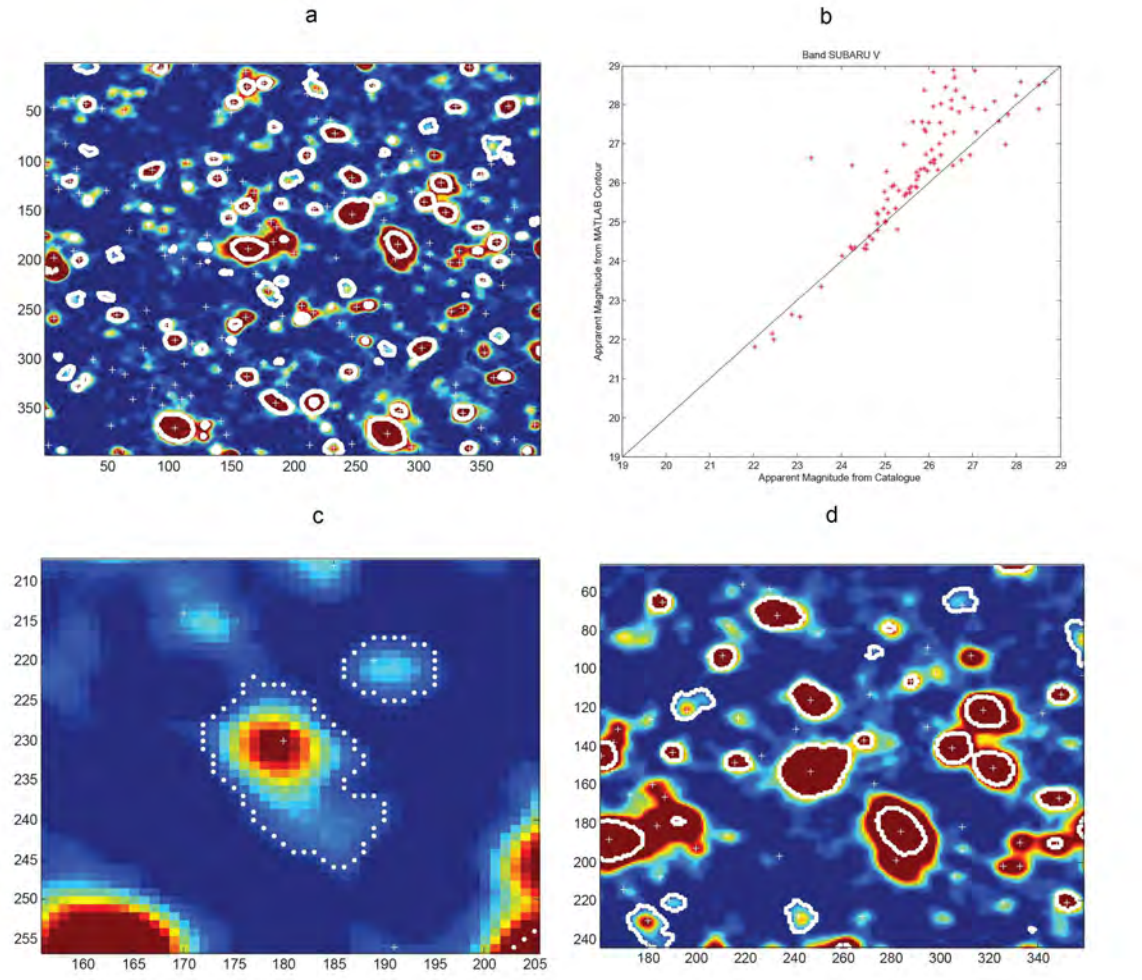

Figure 5.50: 259 COSMOS real sources. Fig a MATLAB contours(white). Fig b. Apparent magnitude from COSMOS catalogue (x axis) and from MATLAB contour algorithm (y axis) . Fig c zoom showing MATLAB contours in detail. Fig d Another zoom showing more MATLAB cotours (white).

resulting in more contours identification.

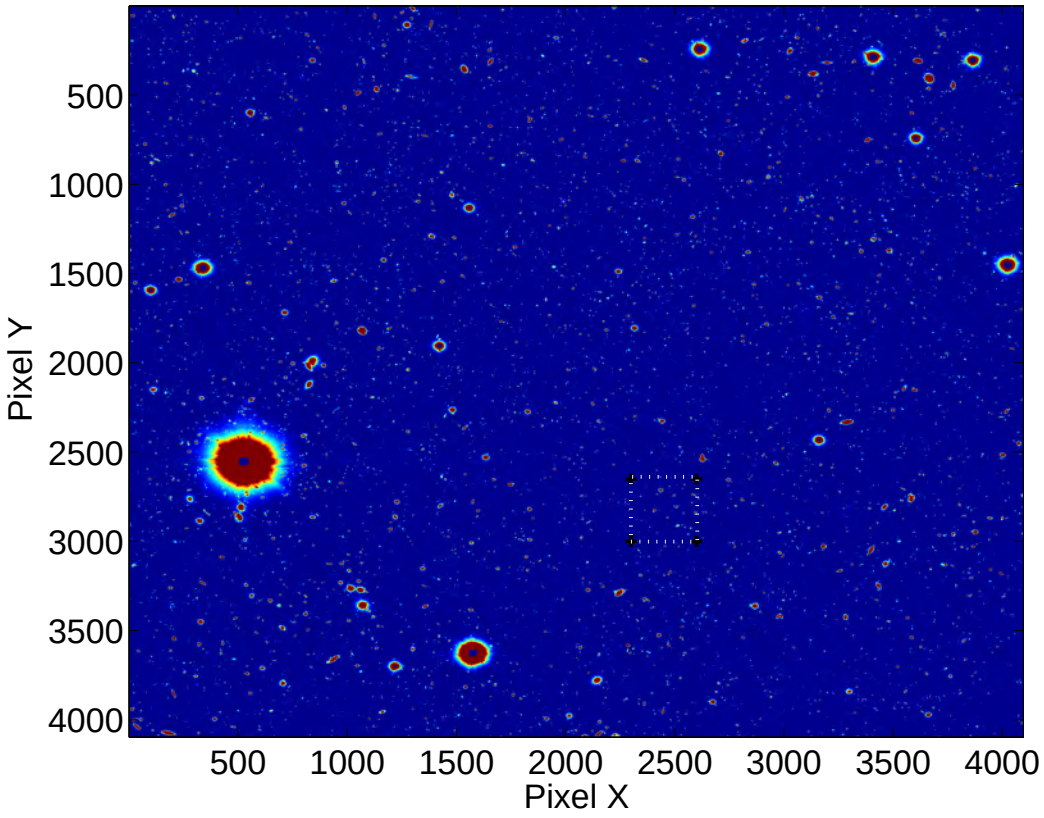

Figure 5.51: Tile 78 of COSMOS Subaru B.

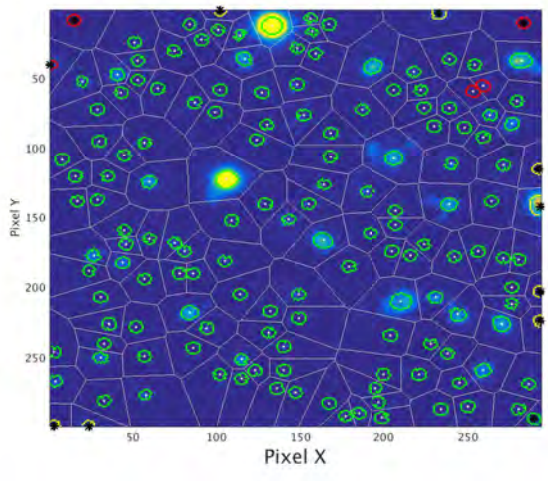

Figure 5.52: Source Labelling for CFHT U.

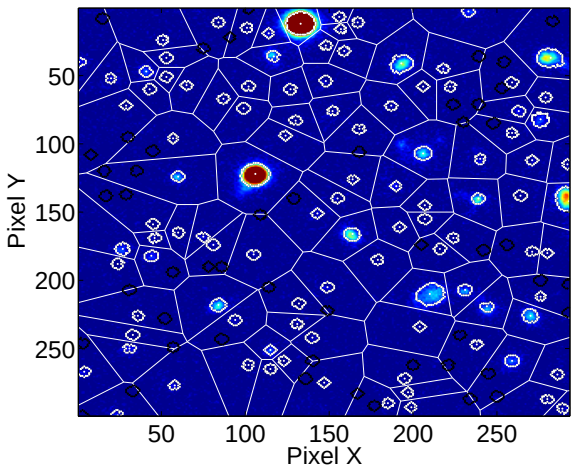

Figure 5.53: GCES contours: Measurement(white). No measurement (black).

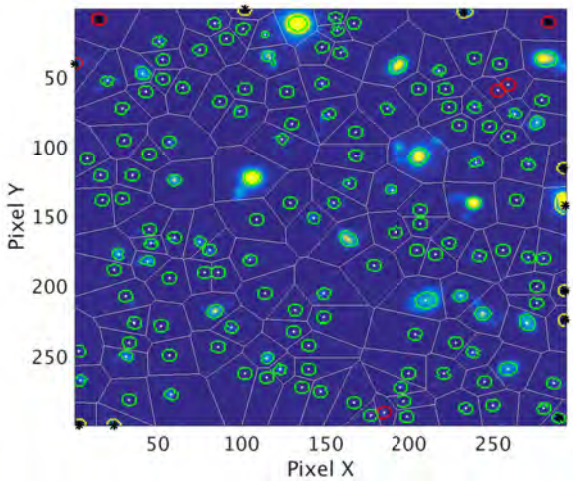

Figure 5.54: Source Labelling for SUBARU B.

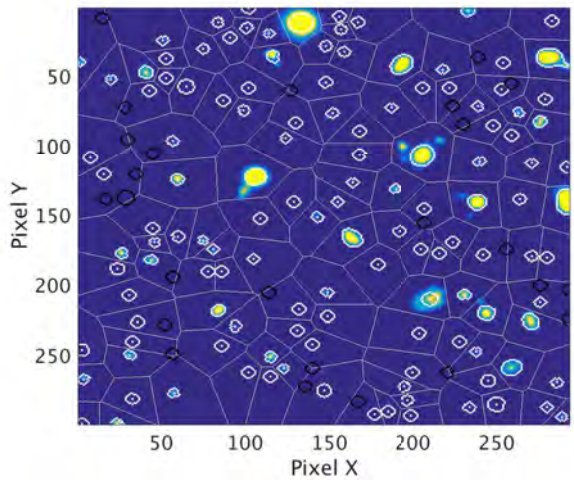

Figure 5.55: GCES contours: Measurement(white). No measurement (black).

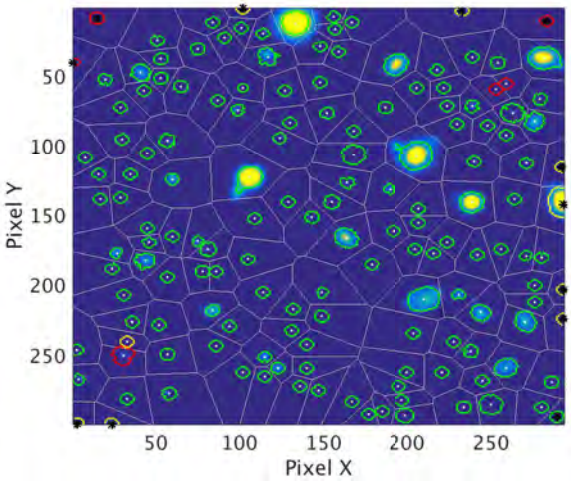

Figure 5.56: Source Labelling for SUBARU V.

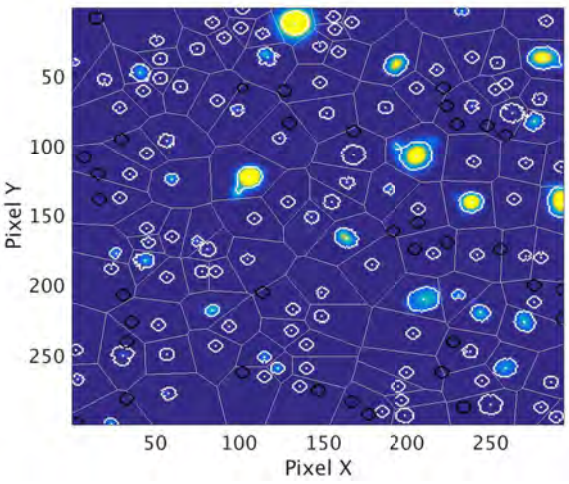

Figure 5.57: GCES contours: Measurement(white). No measurement (black).

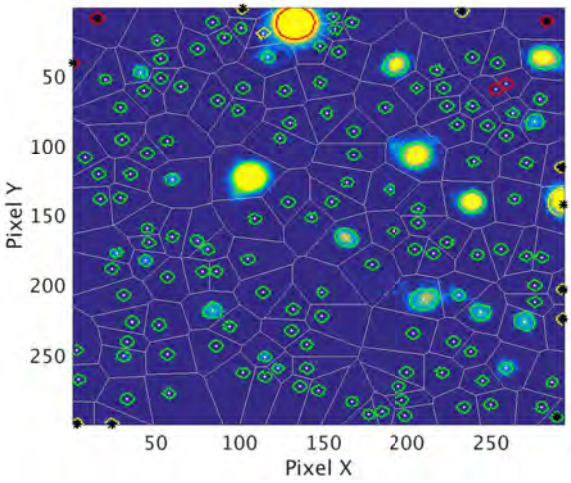

Figure 5.58: Source Labelling for SUBARU G.

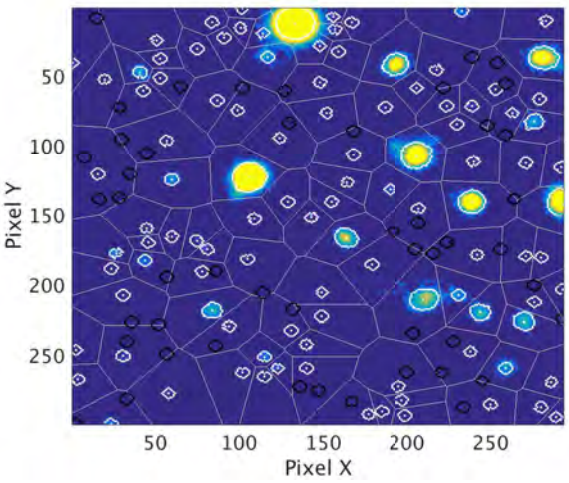

Figure 5.59: GCES contours: Measurement(white). No measurement (black).

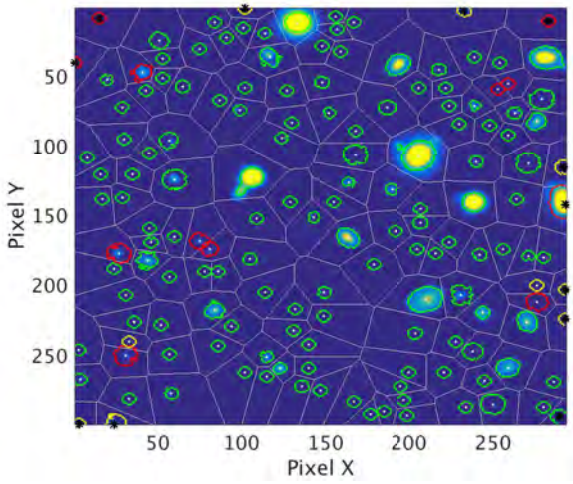

Figure 5.60: Source Labelling for SUBARU R.

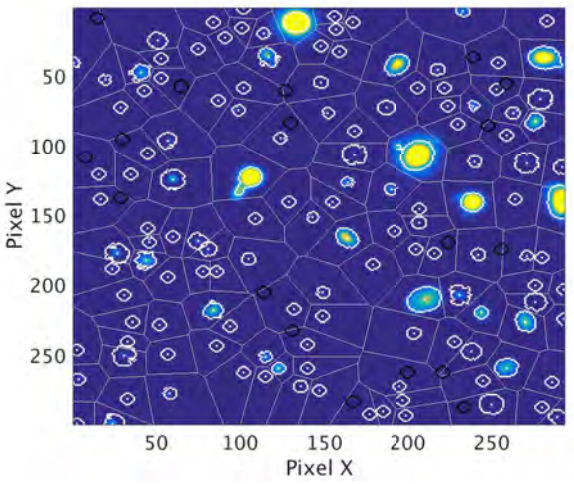

Figure 5.61: GCES contours: Measurement(white). No measurement (black).

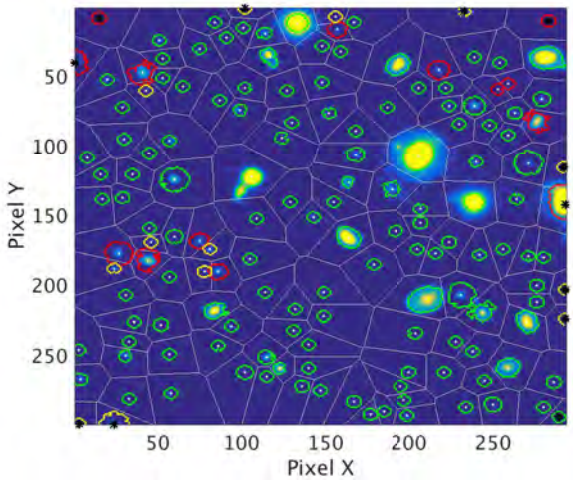

Figure 5.62: Source Labelling for SUBARU I.

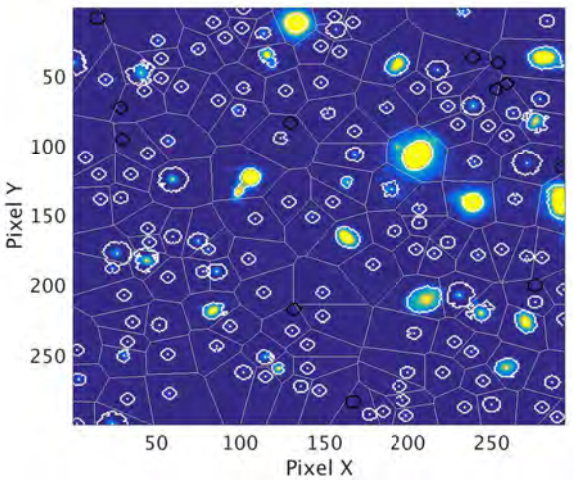

Figure 5.63: GCES contours: Measurement(white). No measurement (black).

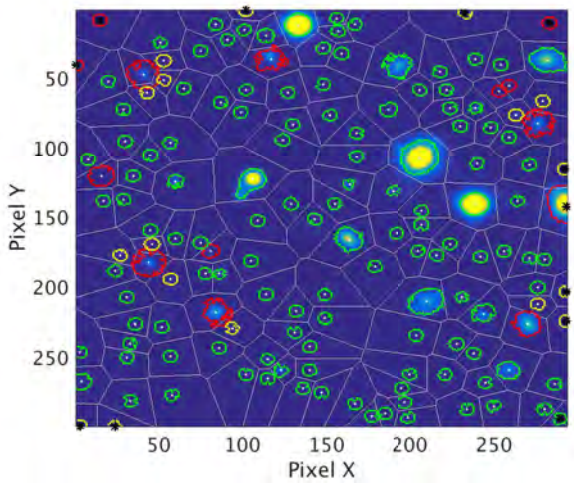

Figure 5.64: Source Labelling for SUBARU Z.

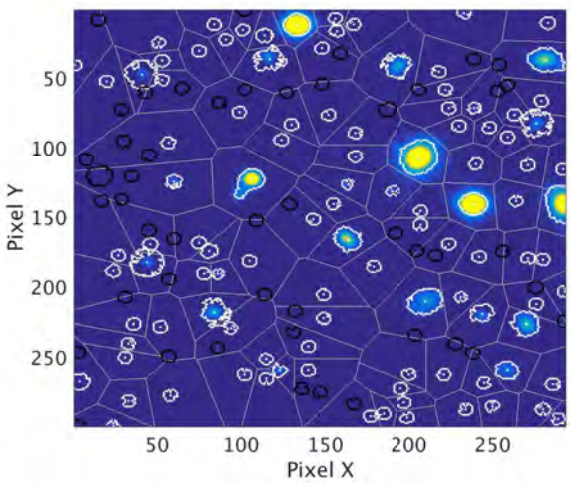

Figure 5.65: GCES contours: Measurement(white). No measurement (black).

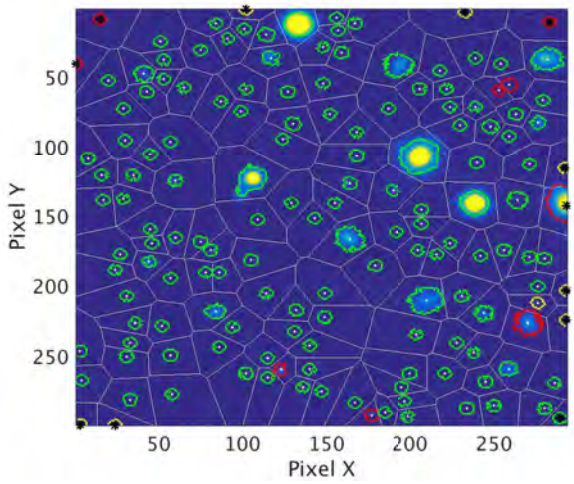

Figure 5.66: Source Labelling for CFHT I.

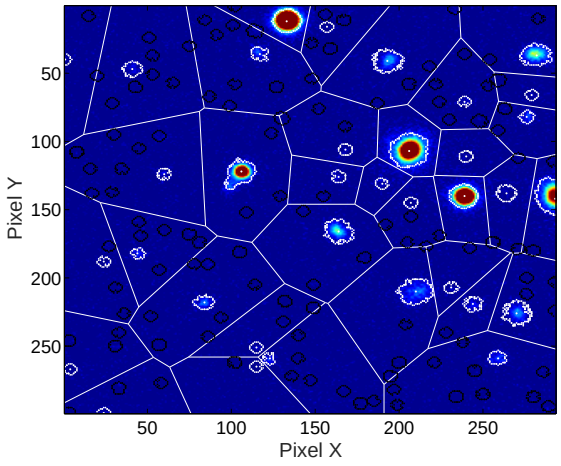

Figure 5.67: GCES contours: Measurement(white). No measurement (black).

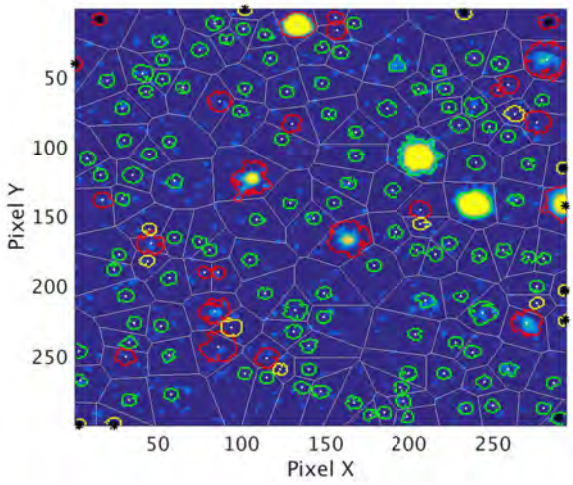

Figure 5.68: Source Labelling for UKIRT J.

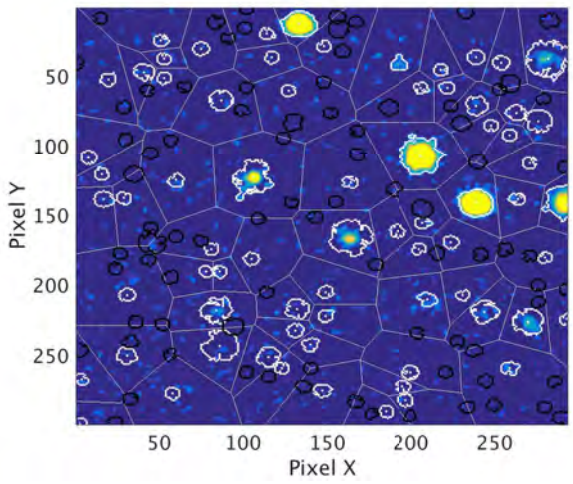

Figure 5.69: GCES contours: Measurement(white). No measurement (black).

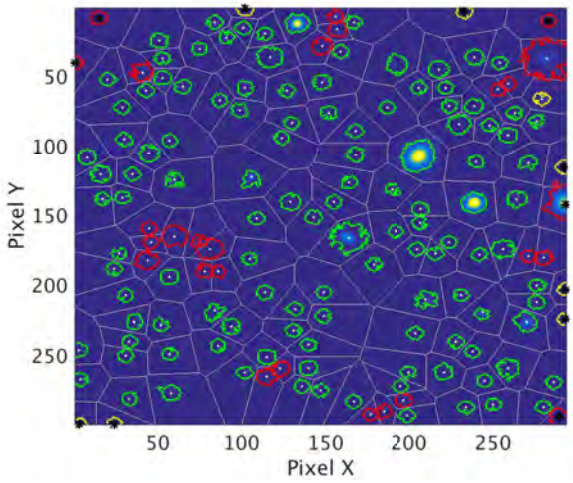

Figure 5.70: Source Labelling for CFHT Ks.

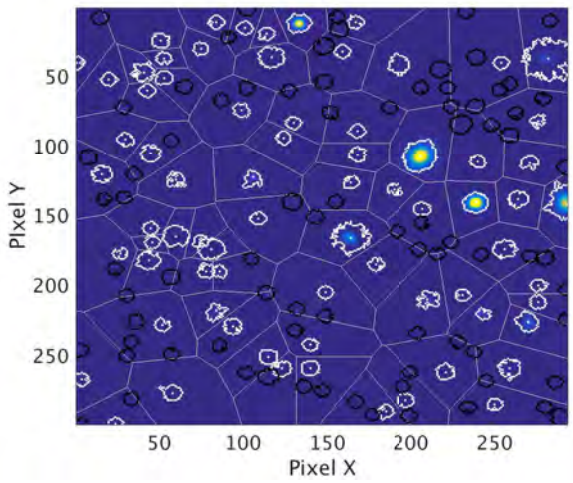

Figure 5.71: GCES contours: Measurement(white). No measurement (black).

5.3.4 Running the GCES with COSMOS Data Set

This section demonstrates the use of the GCES in full with the overall goal of improving the quality of the input data. The main goal of this section is to demonstrate and assess the capability of the GCES when used in full and with the feedback configuration indicated in Figure 5.72. This exercise will allow us to evaluate contributions and limitations of this total GCES usage through the specific use cases stated. Therefore, it is not intended to offer here a statistical representative result because the capability of the full system usage can be demonstrated through specific individual use cases. Each of the modules which composed the GCES have been exhaustively exercised with statistical representative synthetic and real data set as shown in previous and this chapter.
As indicated in Chapter 3, the versatility of the pipeline of this research allows us to use it in a wide spectrum of possibilities. Here we describe the most representative use case scenarios and present the results with real COSMOS data sets along with the connected critical assessment in terms of contributions and limitations from this research.

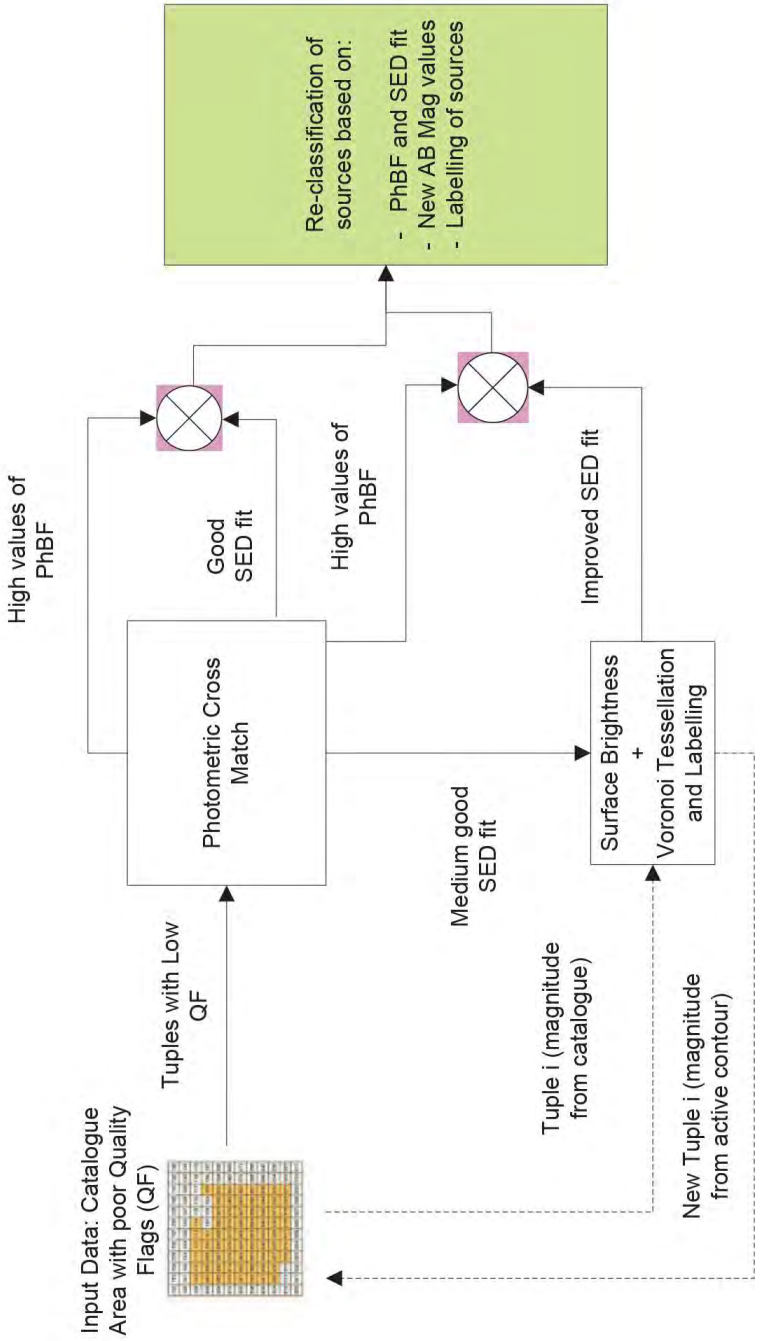

Figure 5.72: Practical execution of the pipeline with COSMOS catalogue

From the block diagram shown in Figure 5.72, we select the following three run cases indicating the contribution of the GCES towards those use cases:

- *Use Case I- From Dubious to Plausible Cases*: in this use case we take COSMOS tuples with Photometric Bayes factor values in the range of uncertain match and we re-run the cross matching with the magnitude values yielded by our active contour approach. Then we evaluate if the initially dubious status of matching for those tuples have been better clarified.
- *Use Case II - Traditional vs Intelligent Aperture Benchmark*: in this use case we compare the traditional Kron ellipses from SExtractor package with the active contour of our research. We evaluate this comparison in a quantitative and qualitative manner.
- *Use Case III - Discovery Capability*: we select COSMOS sources with problems in the magnitude measurement (i.e. 99 or -99 values) and we obtain their magnitudes values yielded by our active contour algorithm. With the new magnitude values we compute the photometric cross matching to determine if it is possible to identify sources not possible by the traditional method. If this is the case, we set up the basis for determining candidates of sources to be better measured in further surveys (maybe with a longer exposure time).

The overall goal from the above use cases is to demonstrate the capability of the GCES in extracting additional information which contributes to refine the confidence level of the data obtained in the COSMOS multi band deep survey in the first place.

Use Case I - From Dubious to Plausible Cases

We ran te GCES in all the COSMOS survey Tiles with low quality flags (columns 41 to 44 of cosmos zphot catalogue) and with a good deep mask (column 45 of this catalogue).

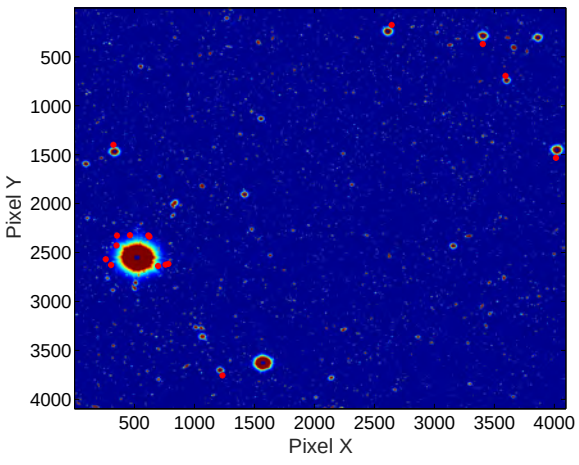

Figure 5.73: Dubious Sources.

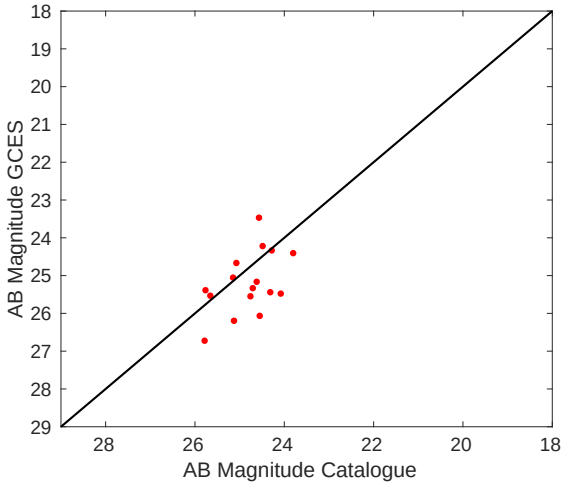

Figure 5.74: GCES versus Catalogue for the dubious sources.

We have conducted a search of COSMOS tuples with low quality flags and looked for those with photometric Bayes factor between 0.05 and 2. This yielded three cases, as indicated in Table 5.1. These are, therefore, good candidates for recomputing their apparent magnitude based on our

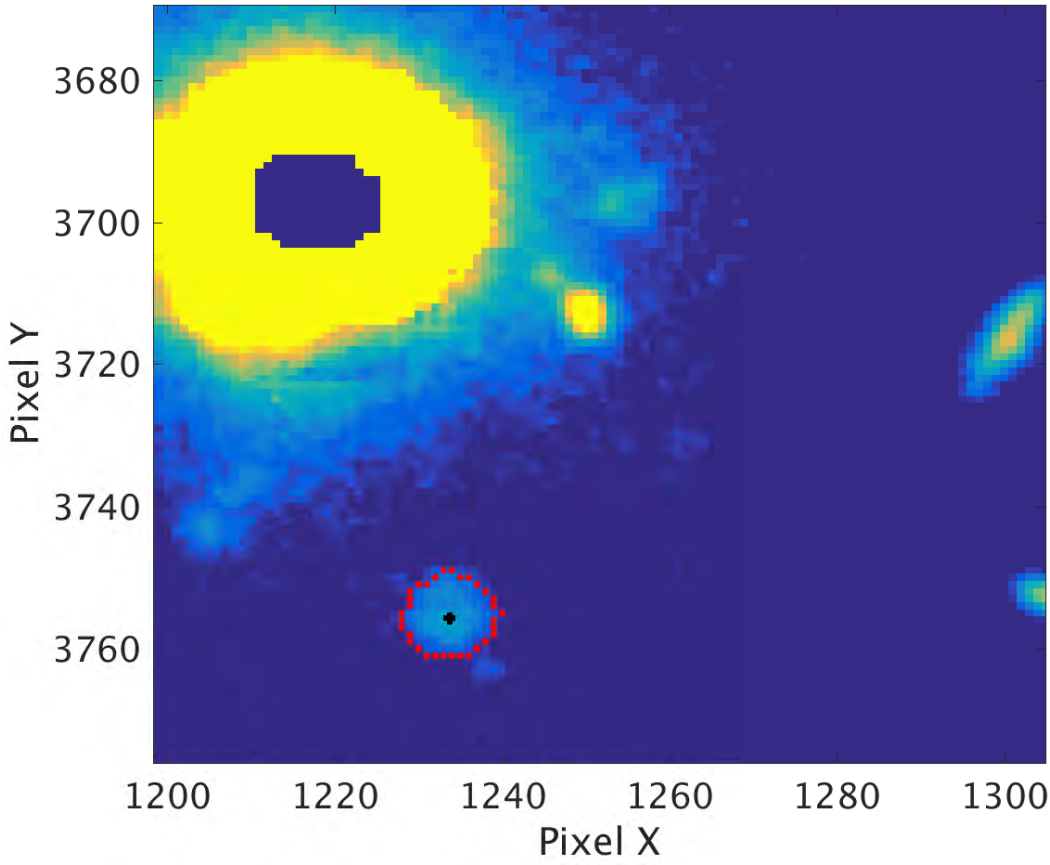

Figure 5.75: Active contour in one dubious source with high PhBF

Table 5.1: Use Case I for three COSMOS sources with PhBF values between 0.05 and 10

Use Case I	Source 1 1248059	Source 2 1248806	Source 3 1256526
Initial PhBF	0.075276	1.192051	0.153596
New PhBF	0.186741	0.0026874	0.00073938

GCES active contour and with these new values, to recompute the photometric Bayes Factor and the relevant best SED fitting. Figures 5.76 to 5.81 show the initial and new best SED fitting results for the three sources selected for this use case.
The main contribution in this area is to improve the discrimination between matches and no matches when confronted with dubious cases by improving the brightness vaues of those bands with low quality flags in the first place.

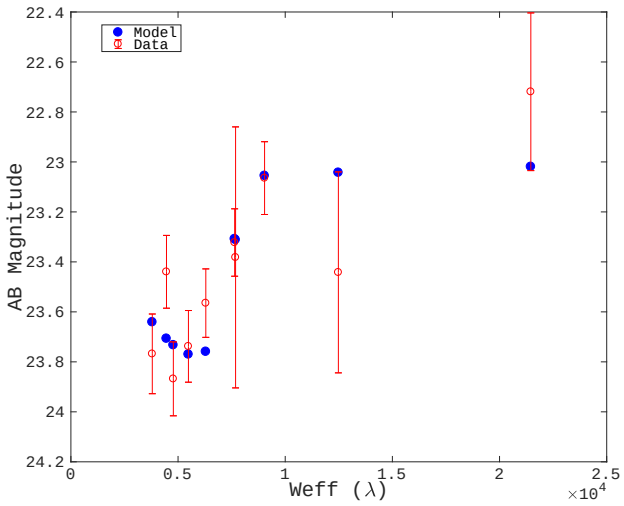

Figure 5.76: Best SED fitting for the COSMOS Source 1248059, before having computed the new PhBF with the active contour AB magnitude for bands CFHT U, SUBARU B, ip and gp. The best SED fitting is reached for a SED spiral SB7_A_0 galaxy and a redshift value of 3.04.

Use Case II - Traditional vs Intelligent Aperture Benchmark

As referred above in various sections, the COSMOS real data considered in this research, in its various releases of 2006, 2008 and later 2015 does not allow us to re-build the contour from which the apparent magnitudes were derived. In addition to that manual steps were performed for the ellaboration of the public catalogues. With all this, it is very difficult to reach a reliable comparative basis with real data in terms of the active contour performance of our GCES versus the catalogues.
However, COSMOS morphology catalogue of 2005 contained information which allow the re-build of Kron ellipses. The values of magnitudes (ISO, APER, AUTO) of this catalogue compared to the photometric catalogue of 2008 contain an important dispersion, not to be underestimated, as it can

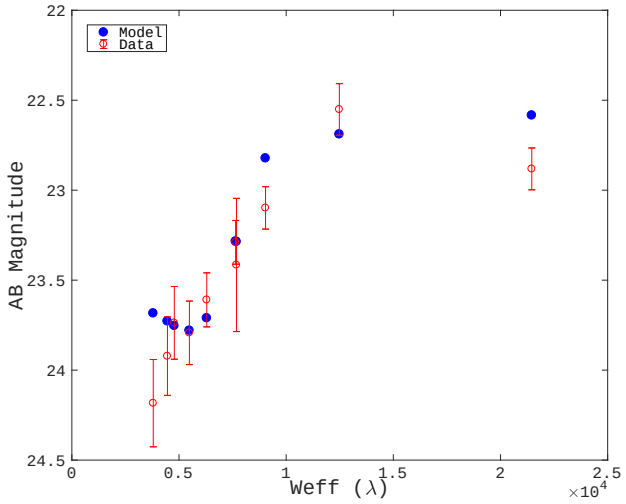

Figure 5.77: Best SED fitting for the COSMOS Source 1248059, after having computed the new PhBF with the active contour AB magnitude for bands CFHT U, SUBARU B, ip and gp. The best SED fitting is reached for a SED spiral SB5_A_0 galaxy and a redshift value of 2.96.

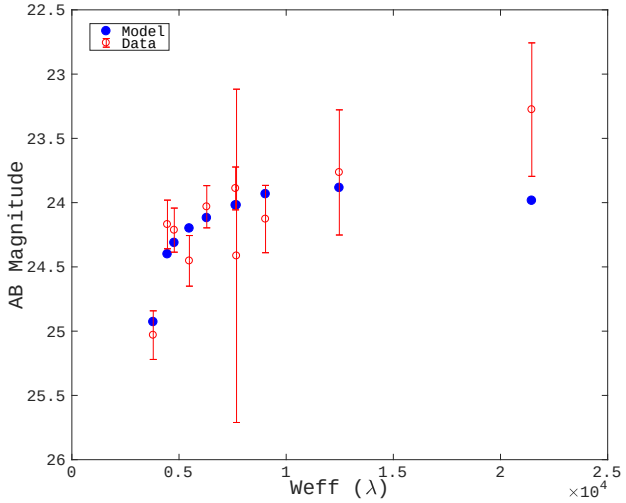

Figure 5.78: Best SED fitting for the COSMOS Source 1248806, before having computed the new PhBF with the active contour AB magnitude for bands CFHT U, SUBARU B, ip and gp. The best SED fitting is reached for a SED spiral SB4_A_0 galaxy and a redshift value of 1.88.

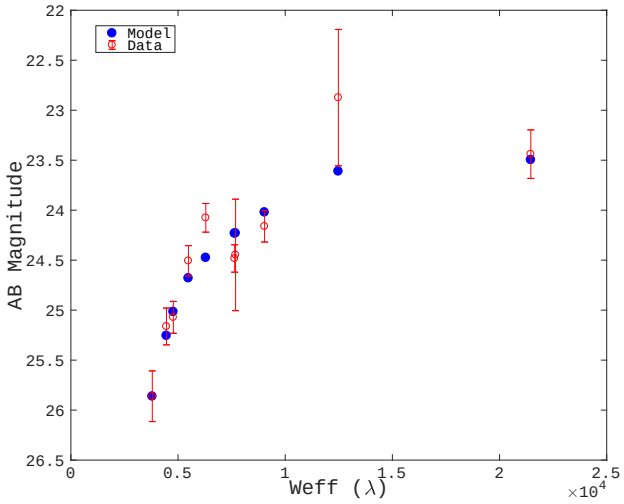

Figure 5.79: Best SED fitting for the COSMOS Source 1248806, after having computed the new PhBF with the active contour AB magnitude for bands CFHT U, SUBARU B, ip and gp. The best SED fitting is reached for a SED spiral Sdm_A_0 galaxy and a redshift value of 1.56.

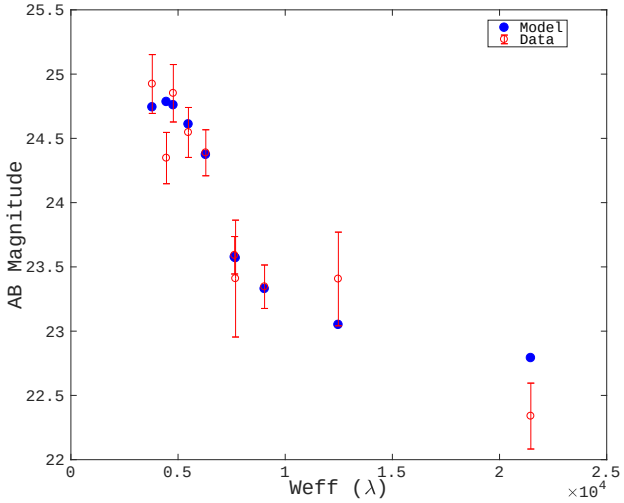

Figure 5.80: Best SED fitting for the COSMOS Source 1256525 before having computed the new PhBF with the active contour AB magnitude for bands CFHT U, SUBARU B, ip and gp. The best SED fitting is reached for a SED spiral SB2_A_0 galaxy and a redshift value of 2.48.

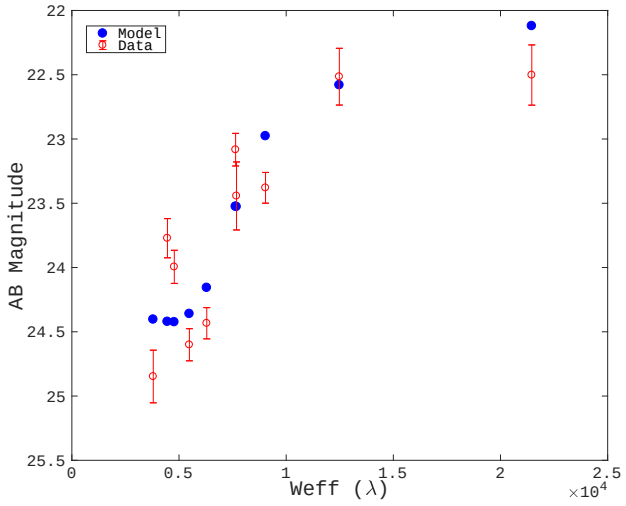

Figure 5.81: Best SED fitting for the COSMOS Source 1256525 after having computed the new PhBF with the active contour AB magnitude for bands CFHT U, SUBARU B, ip and gp. The best SED fitting is reached for a SED spiral SB1_A_0 galaxy and a redshift value of 2.56.

be seen in figure 5.88.
For the sake of having the possibility to observe both Kron ellipses and active contour with COSMOS real data set, we have performed a cross-correlation between COSMOS morphology catalogue (2005) and COSMOS photometry catalogue(2008) Tile 78 and we have selected 170 sources in both catalogues, suitable candidates for this benchmark.
Figures 5.82 to 5.84 shows three cases where differences between Kron and active contour can be observed easily. In Figure 5.82 we observe the case of one source for which the GCES contour identifies better the boundaries of the object. Figure 5.83 shows the case of contouring a faint source. In this case both GCES and Kron show good behaviour in recognizing the boundaries of the object. Figure 5.84 shows the case in which neither Kron nor GCES show good performance in the correct identification of the boundaries of the object. Due to the difficulty of finding a statistical representative basis of the same sources in both morphology and photometry catalogues, the different nature and date of both catalogues and without knowing the steps followed between Kron ellipses and the final apparent magnitude of the catalogues, an exhaustive comparative exercise does not seem to bring added value for the comparative assessment on the performance of contour determination. Therefore the only purpose of the mentioned figures along with figures 5.85 to 5.87 is to present a real case of comparing Kron versus active contour for real sources contour determination and mangitude compuation, bearing in mind the limitations of the catalogues described here.

From Figures 5.85 to 5.88 we show evidences of the important variations in magnitude values from different releases of the COSMOS catalogue for the Subaru B band, as well as from the various measurement methods. Figures 5.85 to 5.87 show a considerable systematic from the diagonal. This means that our GCES system varies considerably in the magnitude values when compared with the COSMOS morphology catalogue for the ISO and AUTO magnitudes. The less dispersion is shown when comparing the GCES magnitude values with the COSMOS morpholgoy APER values, as shown in Figure 5.87.

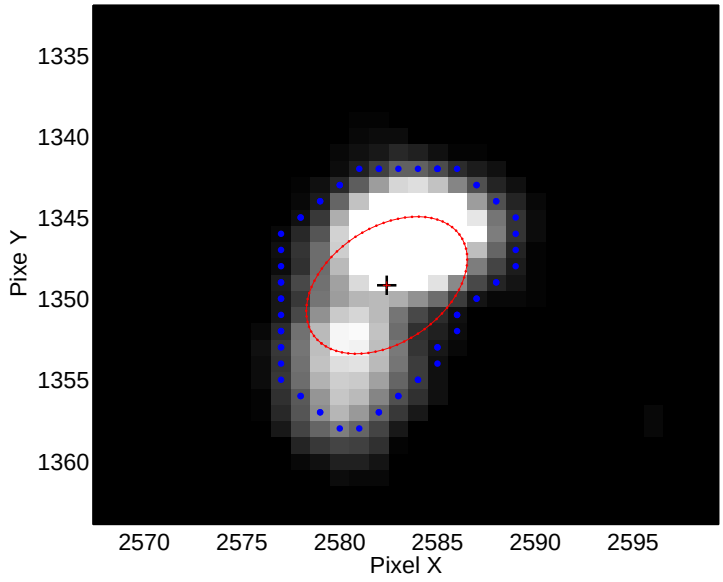

Figure 5.82: Subaru B. GCES vs Kron. Case I.

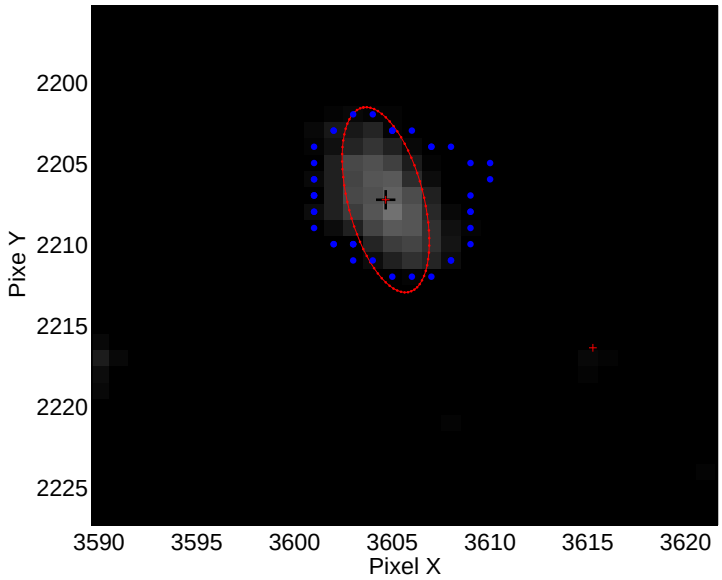

Figure 5.83: Subaru B. GCES vs Kron. Case II.

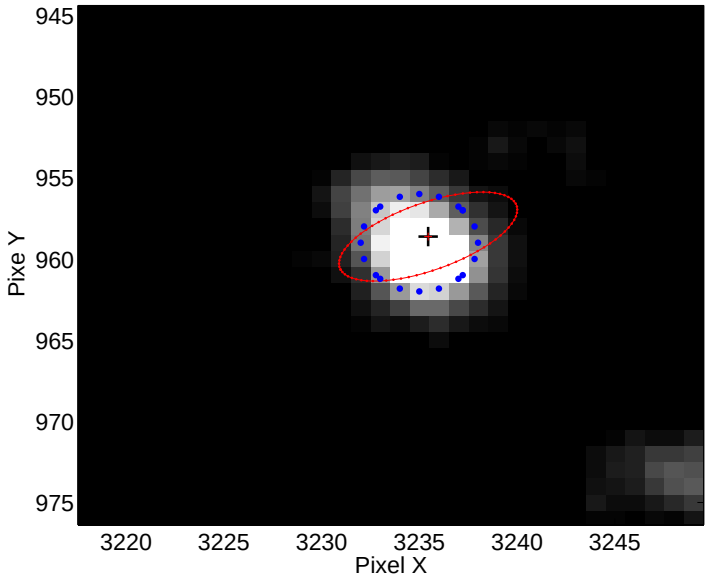

Figure 5.84: Subaru B. GCES vs Kron. Case III.

Figure 5.88 shows the variability among COSMOS catalogues, for the AUTO magnitude values in the Subaru B band. As we know that Subaru B has some calibration problems declared in [Capak et al., 2007], we also represented the comparison for the Subaru i band. In effect, we observed that the scatter of Subaru B is not seen in Subaru I.

Use Case III - Discovery Capability

This use case consists in identifying those sources which originally had the value of 99 or -99 for certain bands. Those tuples containing at least one band with one of these values are normally excluded from any scientific study.Then, the purpose of our use case here is to use our GCES full system on those bands with no valid values and to apply our active contour algorithm. If the result of our active contour yields a valid apparent magnitude value, we can proceed with the Photometric cross matching problem for the tuple and the connected best SED fitting. In this sense we demonstrate that our GCES can contribute to refining the brightness information on sources which initially kept no valid values. With the use of the photometric cross matching in a second step and with the new yielded magnitude values we get confidence that the outcome of our GCES system is good enough to re-ingest those tuples into the study.
To illustrate the use of this case, we selected the source 1246289 with COSMOS catalogue value of 99 for the apparent magnitude in the bands Subaru zp and ip. We applied the use case as described above and the results can be observed in figures 5.89 and 5.90. The new magnitude obtained for both bands, zp and ip were 24.55 and 22.764. One important limitation which can impact on the photometric bayes factor is the selected value of sigma magnitude for the sources with no valid mangitude values. For this case we chose 0.5. The photometric Bayes factor obtained was low so that no conclusive decision on match or no match can be reached. A study on the best approach to select the sigma magnitude value for these cases can be considered in as a future line of research.

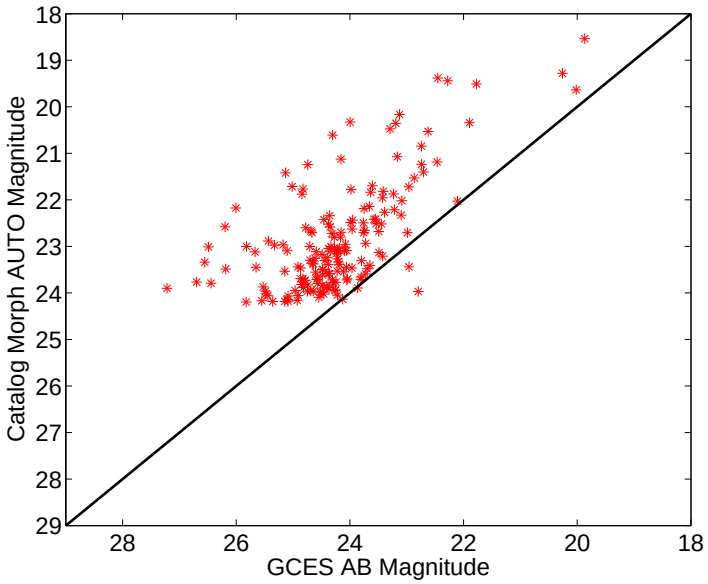

Figure 5.85: Values of apparent magnitudes (AUTO MAG) of Cosmos Morphology catalogue compared to GCES values for the 170 sources.

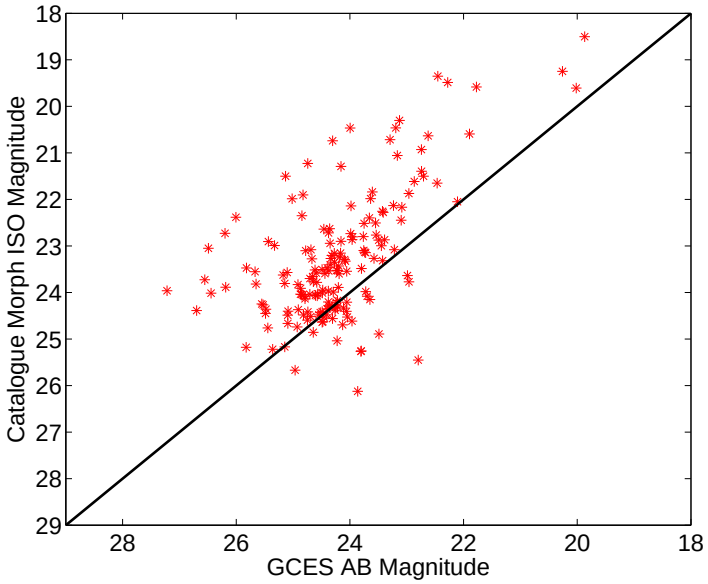

Figure 5.86: Values of apparent magnitudes (ISO MAG)of Cosmos Morphology catalogue compared to GCES values for the 170 sources

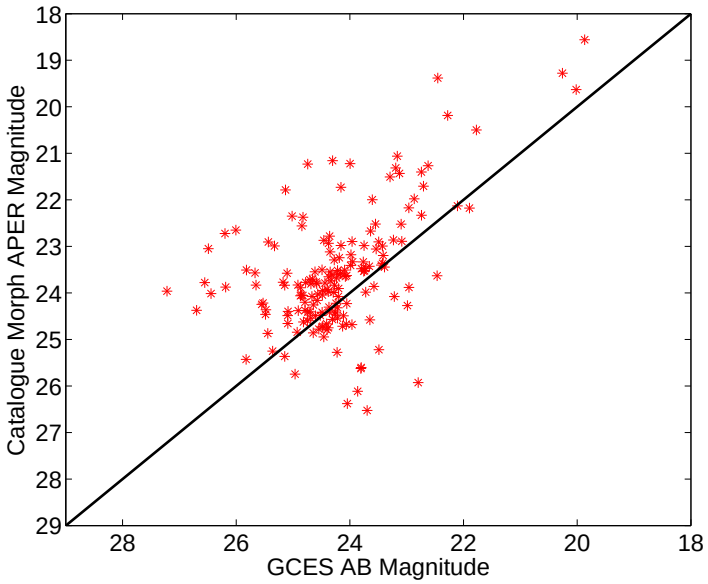

Figure 5.87: Values of apparent magnitudes (APER MAG) of Cosmos Morphology catalogue compared to GCES values for the 170 sources.

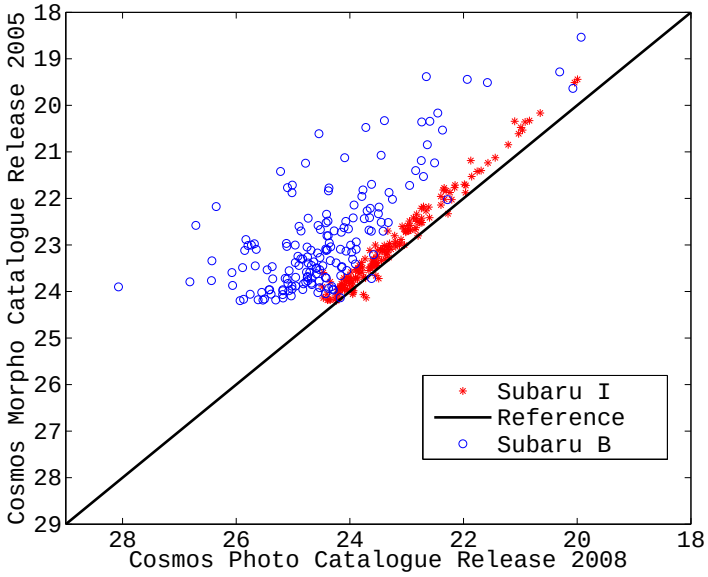

Figure 5.88: Value of apparent magnitude of COSMOS Morphology versus apparent magnitude of Cosmos Photometry catalogue.

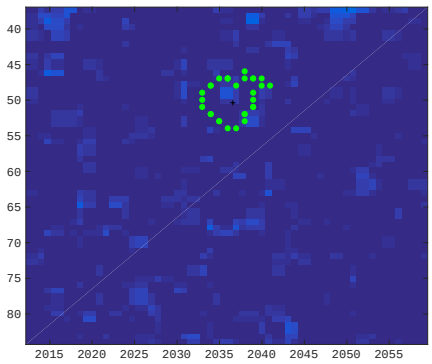

Figure 5.89: Demonstration of Use Case III for the discovery capability of source 1246289 of COSMOS 2008 release catalogue in the zp and ip Subaru bands. Initially the catalogue yielded a value of 99 for both bands. From our GCES we obtained a value of 24.55 for band zp and of 22.764 for band ip.

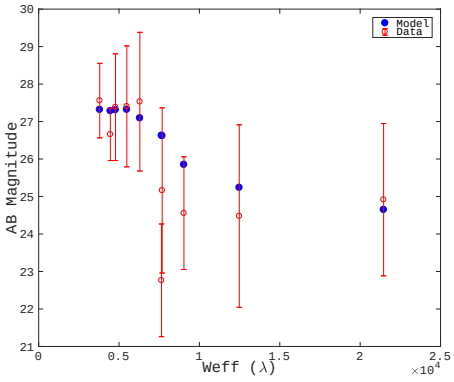

Figure 5.90: Photometric cross matching computation for the 10-band tuple of the source 1246289. With the new AB magnitude for band z and for band ip. The value of the Photometric Bayes Factor is 2.3345 and the Best SED fitting is obtained for SED spiral galaxy SB0_A_0 and redshift of 2.64.

5.4 Critical Assessment

This section presents a critical assessment of the results presented in Section 5.3 and in Sections 4.4 to 4.6. With this we provide a factual evaluation of our Galaxy Cataloguing Expert System in terms of its capabilities as well as its limitations when confronted with synthetic and with real data. This evaluation is divided into three objectives of the thesis, through Sections 5.4.1 to 5.4.3. A dedicated evaluation on the overall GCES is provided in section 5.4.4, and a dedicated assessment on computational aspects of our system is provided in section 5.4.5

5.4.1 Photometric Cross-Matching

We have described the implementation of a Bayesian framework proposed for the problem of the identification and classification of multi-band matched galaxies. We have studied the influence of various sources of uncertainty and various priors leading us to a good understanding of how much and how well can we learn from the model and how good the quality of the photometric Bayes Factor is for the classification of cross-matched versus non cross-matched galaxies.
A sound consistency has been found between true cross-matched sources, high Bayes Factor and good fit with the relevant SED template. Similarly, lack of that consistency is obtained for the case of non cross-matched results. The threshold values of the photometric Bayes factor to indicate when the photometric cross matching yields a conclusive or inconclusive result seems to be linked to the quality of the input data itself. In this sense, a good approach to establish an adequate threshold would be to consider from which value of the photometric Bayes factor the SED fitting is acceptable. A correspondence (not necessarily linear) was observed between the number of bands involved in the

photometric cross matching and the threshold value of Photometric Bayes Factor. It is important to note here that we use the notion of "good fit" to refer to the case of finding a SED function for which the photometric values of the galaxy under consideration are found within the $\pm\sigma$ values of magnitudes for that SED.
The influence of the different priors is considerably small because the overall result in terms of the order of magnitude of the Bayes Factors involved is generally retained. Therefore the robustness of the Bayesian Inference implemented here seems adequate for the purpose presented. However, the range of values of the Bayes Factor varies considerably for the case of the flux prior, being generally larger than the corresponding values for the Uniform prior. Similarly we also observed that the implicit prior based on the 2-way approach compared to the 10-way yields generally lower values of the Bayes Factor for the same cross-matching results.
The use of photometric information as implemented for the multiband matching improves the identification of matches in crowded areas where precision in astrometric positions may be compromised. The SED fitting derived from the implementation proposed here can then resolve the astrometric degeneracy. When faint sources are part of the data under consideration, the combined astrometric and photometric Bayes Factor would allow a further refinement in the identification of matches and non-matches. For example, sources with different redshifts can be identified as astrometric matches, whereas the photometric Bayesian inference described here will discriminate them as a non-match without a good SED fitting.
From an implementation point of view, the validity of the SED template library is key for achieving useful results. It is therefore important to consider a step in validating the SED template to be used. In our case this was not needed because we chose a published release of the COSMOS data as well as for the SED library associated to that data. Another implementation aspect to consider is the computational cost, which depends mainly on the integral dimension and also on the complexity of the prior probability distribution. From the results we can conclude that the simple Uniform prior yields the best result still within acceptable computational cost boundaries (around 4000 tuples can be processed in less than 2 hours with a typical computer).
We observed that for the cases of true photometric matches we can achieve a reliable characterization of the galaxies in terms of their best redshift and template type.
Regarding the overall performance of the system, the capability of differentiating true matches from false ones was demonstrated with the validation done on synthetic and real data. For the implementation with the three priors considered, results show that Uniform and surface brightness priors have better performance than the flux prior, as shown by the true to false positive match ratios. The marginalization of the likelihood over the uncertainties of the scale brightness for the best SED fit done with the flux prior improves considerably the computational cost of the photometric cross matching but it seems to lose relevant prior information. As explained in Chapter 3 and demonstrated through the validation results of Chapters 4 and 5, the consideration of uncertainties in the model should not be underestimated. Before using the photometric cross matching solution, a careful validation step needs to be performed to confirm the adequate range of the Smooth factor for the problem in question. From our results, it can be seen that the larger the Smooth Factor is, the smaller the range of $\log_{10} BF$, which means that the evidence offered by the data is smaller. On the other hand, for smaller Smooth Factors the range of $\log_{10} BF$ will be larger and this means that the data increases its capability to offer useful information in the plausible scenarios presented here. It is therefore important to reach an adequate balanced value for the smooth factor. This value is obtained trying different values and evaluating the results as indicated here.

5.4.2 Source Contour Extraction

We find in the documentation of the COSMOS catalogue releases the description of explicit manual steps requiring human expertise in the field. One of the most relevant information

of any astronomical catalogue is the photometric data, which is heavily dependent on the contour of the objects in the image. SExtractor is usually considered one of the main tools from which contour information is obtained. However, all this process needs to be done by experts and involves important manual steps, for tunning the configuration files. In this sense the GCES presented in this research intends to face the two main challenges here, that is, our GCES is fully automatized and does not require expert knowledge in the field.
We have tried in numerous occasions during the time of this research to retrieve COSMOS catalogue information that would allow us to rebuild the Kron ellipses or equivalent contours used when deriving the AB magnitudes. In order to do that dedicated communications were initiated during the past years with various authors of the COSMOS release paper of 2008 but no conclusive reproduceable outcomes were obtained. Lately a communication was kept with one of the authors of the last COSMOS catalogue release of 2016. However as no satisfactory information was obtained in due time, we concluded that the information was not available. The intended exercise will in any case be pursued at a later time. For the scope of this research we based our assessment of the GCES contour determination on COSMOS data through the comparison of AB magnitude values as presented in Section 5.3.2 and 5.3.4. For the case of synthetic data, contrary to the case of COSMOS data, we have the true reference values of magnitudes against which our GCES contour is validated. This allows us to gain the necessary level of confidence in the evaluation of the performance and limitations of our system.
From the validation results with synthetic dataset in section 4.5, we observe that our GCES contouring approach yields results solidly in line with the true reference, for the case of image without noise and without crowded population of objects, that is, under ideal conditions. Regarding the impact of noise, we observed through the various cases presented in Chapter 4, that for a noise level less than 30%, comparable to the one in the COSMOS images, the GCES contour algorithm behaves robustly, that is, there are no large differences in the outcomes of our GCES contours for a not crowded image with and without noise. However the factor of crowding is not so well handled by the GCES contour algorithm. In that case, the increase in object population implying the appearance of confused areas with blended objects cause an important degradation in the performance of our system. As indicated above, a sound benchmark with the COSMOS contouring algorithm was not possible. Instead, we based our assessment on the comparison of the magnitude values yielded by our system and by the true IRAF reference values, for the case of synthetic data, or by COSMOS catalogues for the case of real data.
The active contour approach seems to yield a more accurate result than the traditional methods for the cases of isolated sources. For non isolated sources, the fixed aperture of COSMOS catalogue contour determination is advantageous with respect of the active contour algorithm as this last one can lead to false results by converging beyond the boundaries of the contaminated source. A future line of work in this direction has been identified. Moreover, we observed that our GCES contouring algorithm achieved a good balance between not reaching an exhaustive boundary of the object and avoiding confused boundaries. Indeed, the typical COSMOS images are overcrowded with many blended sources. However our GCES contour follow considerably well the values of the catalogue. This gives us good confidence to exercise our GCES contouring system with special interest on those cases of the catalogue with dubious quality, i.e. low quality flags or even no magnitude measured in specific bands.
The confidence level in terms of correctness of photometric data beyond 25 AB magnitude is considered poor as stated in [Capak et al., 2007] because for magnitudes fainter than 25 the photometric redshifts begin to behave poorly. This can be a possible explanation why the differences in magnitude values between the expert system and the catalogues increase for magnitudes fainter than 25.
When measuring the depth values in magnitude from our GCES versus the COSMOS catalogues we found that there are differences for the various bands. In [Capak et al., 2007] it is indicated that the depth values vary from band to band. This implies that the measurement and/or the preprocessing

of this measurement is not homogeneous in terms of the sensitivity/capability of detection. The differences between GCES and the COSMOS catalogue can be affected by the various calibration and manual steps particular for each band, as indicated in [Capak et al., 2007]. Moreover, we found that if we do not consider the COSMOS magnitude values of bad quality flags, there is a good alignment between the depth values of our GCES and the COSMOS depth values in magnitudes for those cases with good quality flags.
In terms of colour offset and color-color measurements, there is a general good correspondence between the COSMOS catalogue and the GCES system. This is again considering those values of the COSMOS catalogue with good quality flags.
The typical nature of real survey images and the overall results obtained there for both - traditional and our active contour, highlight the importance of continuing the research of active contouring algorithms for the less favourable set up - i.e. those areas of the sky with crowded blended population of extended sources. In effect, when we compare the results from various release of COSMOS, we observe important differences in various bands and areas of the sky. This consolidates the idea that we need to improve the process of measuring the surface brightness of extended blended sources. More details on the planned future line of work is offered in Chapter 6.

5.4.3 Source Labelling

The source tag labelling results obtained with real and artificial data are consistent with the expectations. This means that the source labelling system works effectively in all cases. This is an obvious expectation from a rule based deterministic system like this one.
This part of the GCES is the most versatile one in terms of being adaptable to any context of problems where this rule needs to be applied. Therefore, the effort to adapt it to a completely different type of data sets should be small.
One important limitation is the degradation of performance accuracy of our GCES labelling system when the dimension of the voronoi grid is near the capability of the computational system. During the process of compiling the results presented in this thesis, we were confronted with an accuracy limitation in the function "INPOLYGON" of MATLAB, used in our rule based system. This limitation is expected to be efficiently mitigated by the implementation of the concept of windows where the rule based system is applied. That window is centred around the point of interest and it has to be selected with enough spatial margin to not cutout part of the area of interest.
Another relevant aspect not to be underestimated is the case of the sources in the boundaries of a Voronoi tessellation. These boundaries can be easily identified because at least one voronoi cell vertex will have the value of infinity. Large astronomical surveys, such as COSMOS, normally compile the images in tiles. The sources in the boundaries of one tile are usually observed in central positions in the neighbouring tiles, in order to avoid losing relevant information. Thus, it would be feasible for our GCES labelling system to trace the sources in the boundaries of one tile in neighbouring tiles. This is not part of the scope of this research but it will be considered as a future line of work.

5.4.4 GCES Results: Quality Flags Refinement

The main overall contribution of our GCES is the refinement of photometric information of data sets. This has been demonstrated through the use cases explained in Section 5.3.4.
From the results obtained in the COSMOS grey area, it is confirmed that the photometric Bayes Factor contributes to the refinement of the astrometric identification of the same source across multiple bands. Obviously, this conclusion is also applicable to the non grey area of any catalogue, although it is expected to be specially useful in areas of the catalogue with low quality flags. Derived from this, we can state that the photometric cross matching offers a reliable and sound frame to refine the quality of multi-band cross-match within the catalogue and among different catalogues.

The results obtained from the GCES contouring algorithm for the cases where the COSMOS catalogue yields a magnitude value of 99 or −99 meaning no measurement, confirms the added value of this new approach of automatic source contour extraction. Obviously this new contour information has to be understood as one input to the other elements of information of our GCES as reflected in 5.72.
Finally, the GCES labelling system, used in conjunction with the other elements of the system is a powerful tool for the discrimination of dubious values in terms of matches and of magnitudes.
In addition to the above and considering the versatility of the GCES architecture and its intended use, one can derive new quality flags and build with them an heuristic system to optimize the task of analysing and classifying the astronomical information. This, which is a step further from the basis of this thesis is obviously out of the scope of this research.

5.4.5 Computational Load

The computational cost required for the run of each module of the GCES as well as the full GCES should not be underestimated. The compilation of results in this research has presented important challenges. The less expensive module of GCES is the photometric cross matching and the most expensive one is our active contour algorihm. The source labelling module is not very expensive but it runs into performance problems relatively quickly (e.g. one tile of COSMOS shows small performance limitations).
Initially, a remote server from Johns Hopkins University with up to 4 cores and massive storage capabilities was used allowing an efficient execution of the photometric cross matching software. For the active contour software, a cluster of 10 to 15 MATLAB instances with 12 parallel workers was used from another technical computing environment multiplying, by a factor of 10 the speed of the compilation of the final results as compared to the use of a single machine with only one MATLAB instance.
Therefore, for this research we have not used a dedicated system, but a shared cluster of Matlab R2013a and R2016b 64 bits linux based and a storage quota of around 50 Gb (courtesy of EUMETSAT). The runs were executed in intermittent mode, mainly at night in a low speed profile. The main reason for that is that this system is not intended to be used for this research and no other platform with the necessary requirement was available at UNED at the time of compiling the results. To provide some facts, the average time to process around 100 sources contouring and labelling would be around 2 hours. To compute the photometric cross matching of around 3000 sources around 2.5 hours are required (average with the usage considerations indicated above). And to compute active contour of around 50 sources an average of 50 minutes are required.
In terms of storage, around 2.5 TB of information were stored (considering also MATLAB workspace to reproduce results) in local redundant hard disks.
The MATLAB function tic() and toc() have been used for the computation of the time spent in the modules of our GCES.
It is important to note here that the execution of the full GCES on the platform described would imply a too expensive computational cost when ran on the full COSMOS data survey. Therefore, a subset of the COSMOS data was selected such that it retained the necessary representativeness for the connected assessment on the results. An exception of this was the use of the full COSMOS photometric data set with low quality flags for the problem of the photometric cross-matching and its derived SED fitting. We achieved with this a global refinement of the catalogue cross-matching for those cases of low quality flags indicated in the official COSMOS catalogue.
Considering the various steps used within the overall software code of this research, the Voronoi-cell based search across images and the active contour algorithm are the most expensive ones from a computational cost point of view. The main reason for that is the fact that the structure of this

module is composed of various nested exhaustive search loop of the voronoi cell matrix. In order to optimize this computational cost we used the MATLAB parallel loop with 12 parallel workers for the outer loop and with that we observed a considerable increase of the computational speed of around 50% (2.4 hours now compared to around 4.5 before).
Figure 5.91 shows, for general illustrative purposes, the overall reading of CPU consumption of one of the MATLAB cluster allocated to the HPC1 node, while running our active contour software intensively. Figure 5.92 shows the CPU load allocated to HPC1 while running the two most expensive GCES modules (active contour and ruled-based system across voronoi cells). Similarly Figures 5.93 and 5.94 show the overall CPU use and load for the other node used HPC2. It is important to note here that all these readings are referred to a shared frame where more than one user may be interacting with the system. For the figures presented here we selected a day / time with minimum population of users.

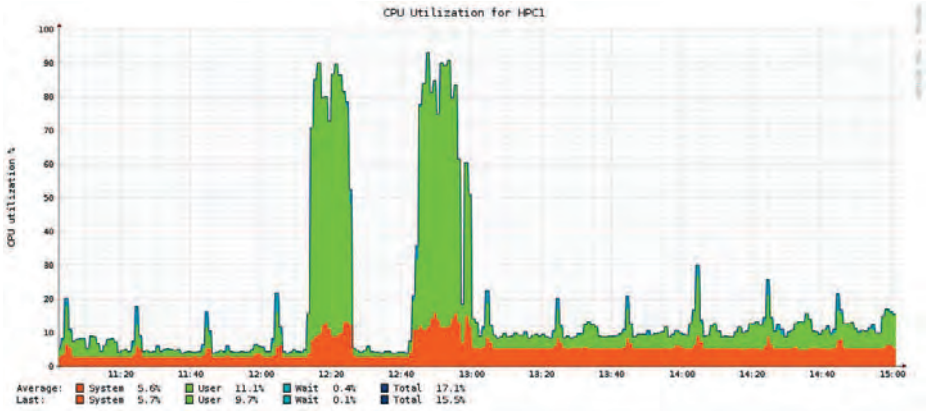

Figure 5.91: Overview of the HPC1 cluster in terms of CPU usage by the user community and by the system.

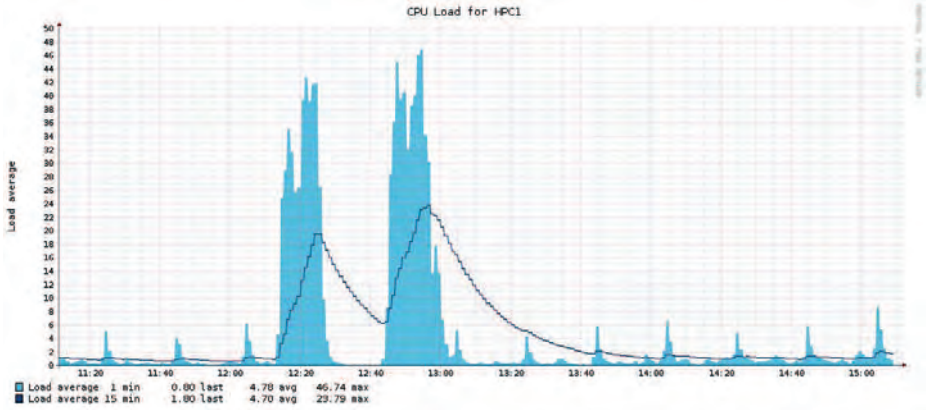

Figure 5.92: Overview of the CPU load for the HPC1 cluster.

A dedicated Linux script was run in order to record the % of CPU and of RAM consumed by the GCES during 2.81 hours while executing the active contour and labelling modules of the GCES in the two cores HPC1 and HPC2 with 12 parallel workers in each of them. Figure 5.95

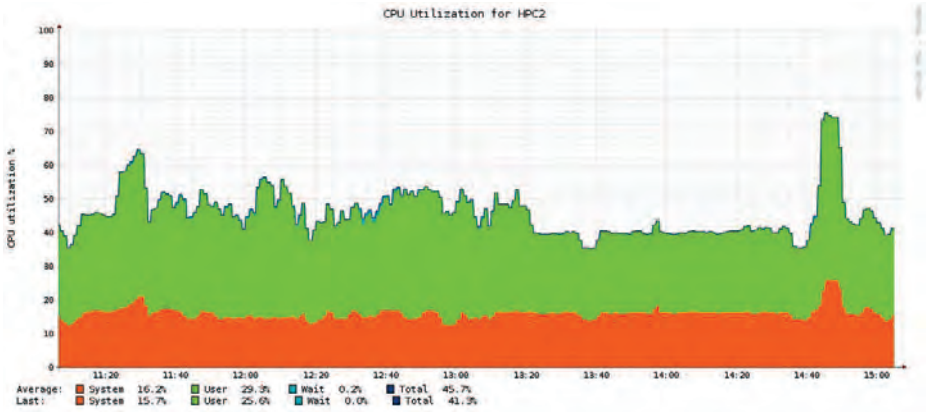

Figure 5.93: Overview of the HPC2 cluster in terms of CPU usage by the user community and by the system.

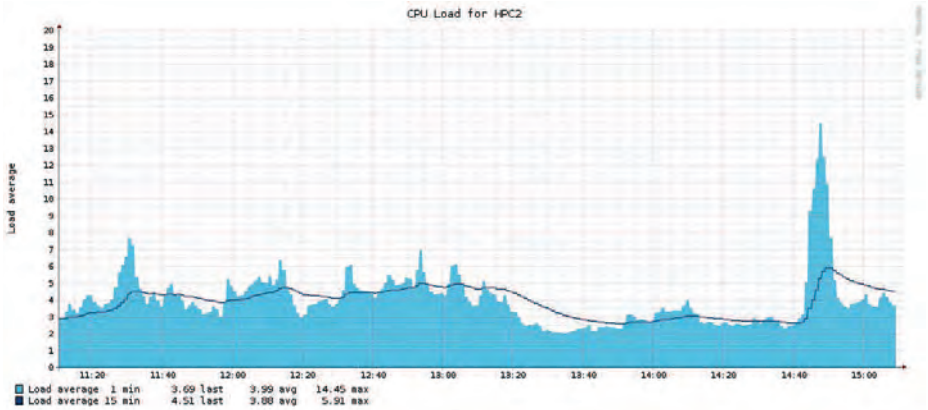

Figure 5.94: Overview of the CPU load for the HPC2 cluster.

shows the overall result during that time. We observe that the process of RAM transfer is constant as expected from an optmized process. Moreover, the CPU consumption show the expected peak for the areas of the code with more computational cost, but it always return to low values. This confirms that there is no computational limitation in the software coded. Figure 5.96 shows a zoom of Figure 5.95 during a short period of time.

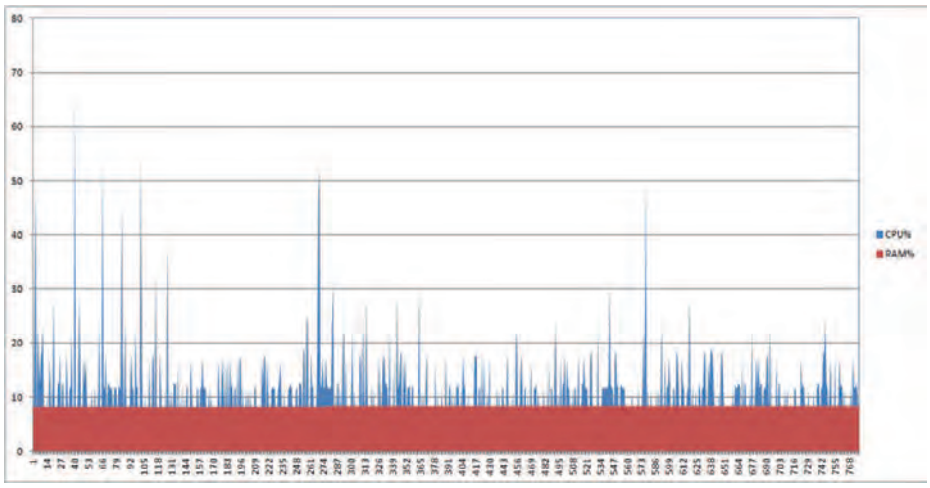

Figure 5.95: Overview of the CPU and RAM load in our user quote during the execution of computational cost modules.

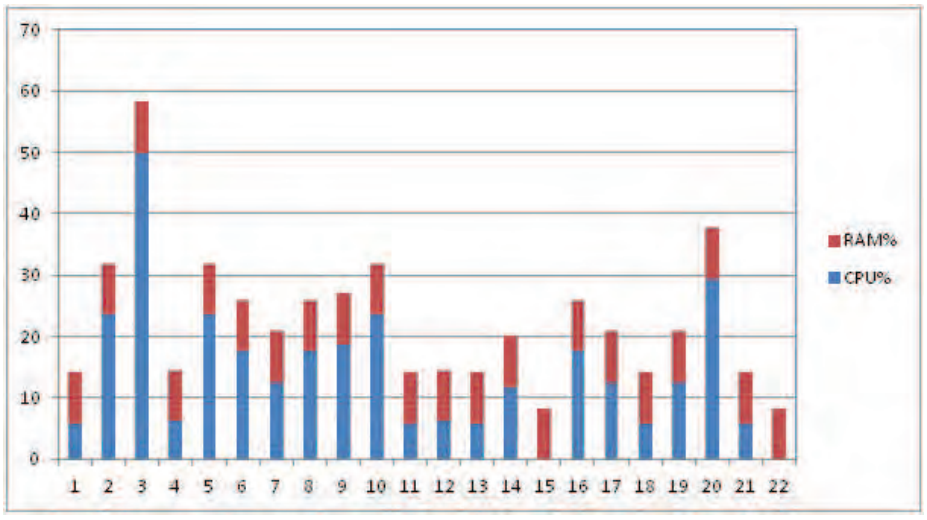

Figure 5.96: Zoom of the CPU and RAM load during a short period of time.

Chapter 6

Conclusions

We summarize here the conclusions derived from modules of our GCES system. In addition to these specific conclusions, the following ones can be derived from the consideration of the full GCES system:

- The implementation of the full GCES on low quality areas of a catalogue improves its overall quality.
- The approach of continual optimization of our system is achieved through the utilisation of the outcomes of one block as input to another block.
- The flexible architecture of our system allows us to use it partially or totally in single, multiple or grouped-band data set configurations.
- Overall the GCES system presented here has demonstrated promising capabilities in offering a complementary source of information to improve the information extracted from the astronomical surveys.

6.1 Photometric Cross-Matching

We have described the implementation of a Bayesian framework prototype which we used to study how much the photometric measurements can contribute to the problem of identifying and classifying multi-band matched galaxies. This is mostly relevant where the astrometry is not decisive. In principle, we can create multiple matched catalogues based on just astrometry or based on both astrometry and photometry, which can become particularly interesting when the results of cross-matching based only on astrometric data are too poor to yield robust conclusions, this leading to missing or incorrect matches.
For the prototype created here we studied the influence of different sources of uncertainty and different priors, leading to a good understanding of how much we can learn from the data, how good the model is, and with that, how good the quality of the photometric Bayes factor is for classifying cross-matched versus non-cross-matched galaxies. We reached the following main conclusions from this study:

- A sound consistency was found between cross-matched sources, high Bayes factor, and good fitting with the relevant SED template.

- The photometric cross-matching prototype was fully validated with synthetic and real data, showing in that respect a very good alignment with the expected known results.
- The photometric cross-matching Bayesian framework implemented on astrometric matches can increase the level of confidence regarding cross-match results of multi-wavelength tuples, which has a particular importance when the astrometric Bayes factor cannot help in determining whether there is a match or not.
- When only astrometry was used to identify matches, the photometric cross-match approach performed a refined discrimination of matches from no matches. An astrometric match where photometry is not considered at all may lead to identifying galaxies with different values in terms of redshift and/or shapes as the same source. Therefore using photometric information as implemented in this thesis improves identification of matches in crowded areas where precision in the astrometric positions may be compromised. The SED-fitting derived from the implementation proposed here can then resolve astrometric degeneracy.
  When faint sources are part of the data under consideration, the combined astrometric and photometric Bayes factor allows a further refinement in the identification of true and spurious matches. Cases, such as overlapping sources that belong to different redshifts, can be identified as astrometric matches, whereas the photometric Bayesian inference described here will discriminate this as a no match with poor SED-fitting.
- In relation to the way in which the different priors affect the model decision problem, we observed that the uniform prior, where no previous knowledge is assumed, retains the best results in terms of discrimination between true and spurious matches. On the other hand, knowledge introduced by the prior based on flux does not always yield the best model decision results in this case. We conclude that the data used here to compute the flux prior might create a slight bias in the overall computation; however, this deserves a further detailed study. In conclusion, the influence of the different priors is very weak because the order of magnitude of the Bayes factor is generally retained. Therefore we consider that the robustness of the approach presented here is supported by the fact that different priors do not substantially change the results.
- From an implementation point of view, the correctness and validity of the SED template library is the key to achieving reliable and useful results. It is therefore important to consider a step in validating the SED template to be used as described here.
- The correctness and validity of the uncertainty described here is a complex topic of current active research. In this respect, the values used in this prototype, being validated through artificial and real data sets, has proven to be adequate (see discussion in Section 5.3.4.
- Another implementation aspect to consider is the computational cost, which mainly depends on the integral dimension and also on the complexity of the prior probability distribution. We therefore recommend studying the capabilities of the computational system being used in order to avoid computational cost underestimation.

6.2 Source Contour Extraction

Based on the results from synthetic and real data of the implementation of the active contour and surface brightness computation described in this thesis, the approach seems to have reached a promising basis for a robust automatic source definition when the source is isolated. The main conclusions of this part of the research can be summarised as follows:

- When comparing magnitudes values from the IRAF synthetic images with the ones yielded by our GCES we observe that our GCES system cope well with realistic level of noise. This realistic noise value is obtained from the values of gain and noise recommended by IRAF and based on typical instrument performance. Therefore, the active contour algorithm proposed can be applied in general with confidence, as long as it is retained the notion that the contours will very probably be incorrect for the blended areas and if the noise of the image is very high.
- When comparing the GCES active contours with the SExtractor equivalent ones (only possible for a few galaxies from the 2005 COSMOS morphology catalogue), we observe promising results for our GCES contour under the conditions referred above. However, due to the few examples available this result can only be judged as preliminary and it is not statistically representative to draw a solid conclusion on how good GCES automatic contouring is compared to the traditional SExtractor contour. Thus, when analysing the Cosmos Morphology catalogue versus the active contour ones, it was observed that the SExtractor Kron ellipses tend to go in some cases beyond the area of pixels which are visually covering the source and in some other cases the main axis of the Kron ellipse is going in a different direction from the one visually apparent of the source itself, whereas the active contour tends to cover the extension of the source more accurately in terms of the main axis of the object, mainly for isolated faint sources, although sometimes its coverage keeps some source pixels outside of the contour (normally these are low brightness pixels in the proximity of the contour frontier). This assessment was carried out using direct visual inspection on a few available sources.
- One relevant difference between our GCES contouring algorithm and the traditional SExtractor Kron ellipses, is the fact that the K factor of the Kron ellipses is specific for each image, whereas our GCES does not need manual tuning of the parameters. This means that in terms of automatization the GCES contouring algorithm represents an improvement compared to the traditional approach.

6.3 Source Labelling

The following list contains the most relevant conclusions related to Voronoi tessellation and galaxy classification:

- The Voronoi tessellation and the simple rule-based system defined in this research have been run in many various sources always showing a correct result in the classification. Therefore the robustness of this approach has been demonstrated.
- As a general conclusion it can be said that the automatic classification of sources as isolated or contaminated has proven very useful in terms of complementary information about other quality flags.
- For the source labelling algorithm, it is important to not underestimate the performance degradation when the Voronoi grid dimension is to large for the computational system capability. In such cases, it is recommended to split the area of study in sub windows bearing in mind the importance of well identifying the sources in the frontiers.
- The management of the boundaries is achieved in our GCES labelling system by identifying those objects whose Voroni cells belong to the boundaries of the tessellation.

6.4 Use of the GCES

In terms of the practicalities associated to the implementation of the pipeline described in this research, the following main conclusions are presented:

- The computational cost associated with the development of the main blocks of this optimisation pipeline needs to be seriously considered. The usage profile for this research seems to have entered into the portal of Big Data for the size of compiled results and for the computational processing set up needed (multiple machines in a share-quota context and 12 parallel workers in the processing chain)
  The computational speed has been found improvable in terms of using this system as it is now and in the current architecture with massive data, as it would be a typical real usage. The notion of a continuous dedicated cluster of machines multiplying the instances of GCES runs could increase the speed to the desired level according to the requirements. Therefore we can conclude that the GCES architecture does not have internal computational limits derived form its design and it is therefore a scalable software.
- Despite the performance disadvantage of MATLAB in the face of other development platforms, it has been a good choice to allow rapid prototyping of the pipeline, while the design of the architecture has been kept in UML form. In this way, the code of the pipeline can be imported into another more suitable development platform.

6.5 Summary of Future Lines of Work

The following summarises a few lines of potential future work as an extension of the current lines of research elaborated here:

- The Voronoi tessellation opens up interesting continuation paths of research, in view of the upcoming massive data with challenging images. One of these possible paths would consist of decreasing the computational cost of the overall processing by splitting the data into groups of Voronoi cells. A second interesting path consists in using the structure of the voronoi tessellations to improve the accuracy of the contour determination, specially in the blended areas of the image.
- The improvement of the active contour algorithm for the determination of boundaries in blended areas has started to be explored, offering promising results, as shown in figure1 6.1 to 6.3.
- The use of luminosity functions instead of apparent magnitude in the photometric cross-matching problems is an interesting path that should be followed.
- The vectorisation and parallel processing capabilities techniques seem suitable features to be explored by developing an upgrade of our GCES baseline presented here.
- The consideration of a heuristic decision automatic system based on the addition of quality flags as explained below seems to be a promising future line of work.

Active Contour in Confusion Areas The areas of confusion are one of the most challenging problems in the computation of the surface brightness of each astronomical source. One preliminary approach explored here consisted of determining the frontier between two or more blended galaxies in the image by computing what we have termed in this research the **brightness radio-vector**. This vector is built up for each source as follows:

- The origin is at the center of the source

- The end is on the boundary of the source
- The vector magnitude is the difference in brightness between the end of the vector and the origin

The brightness radio-vector of each pixel in the confused area will be obtained and those vectors with the larger magnitude will determine the frontier between blended sources.
The following figures 6.2 to 6.3 show examples of the implementation of the approach proposed for blended sources in areas of high overlapping density in the images. The frontier computation of the overlapping areas is obtained by implementing the concept of brightness radio-vector, as described above.

Figure 6.1 presents a preliminary results in the future line of work related to determining the

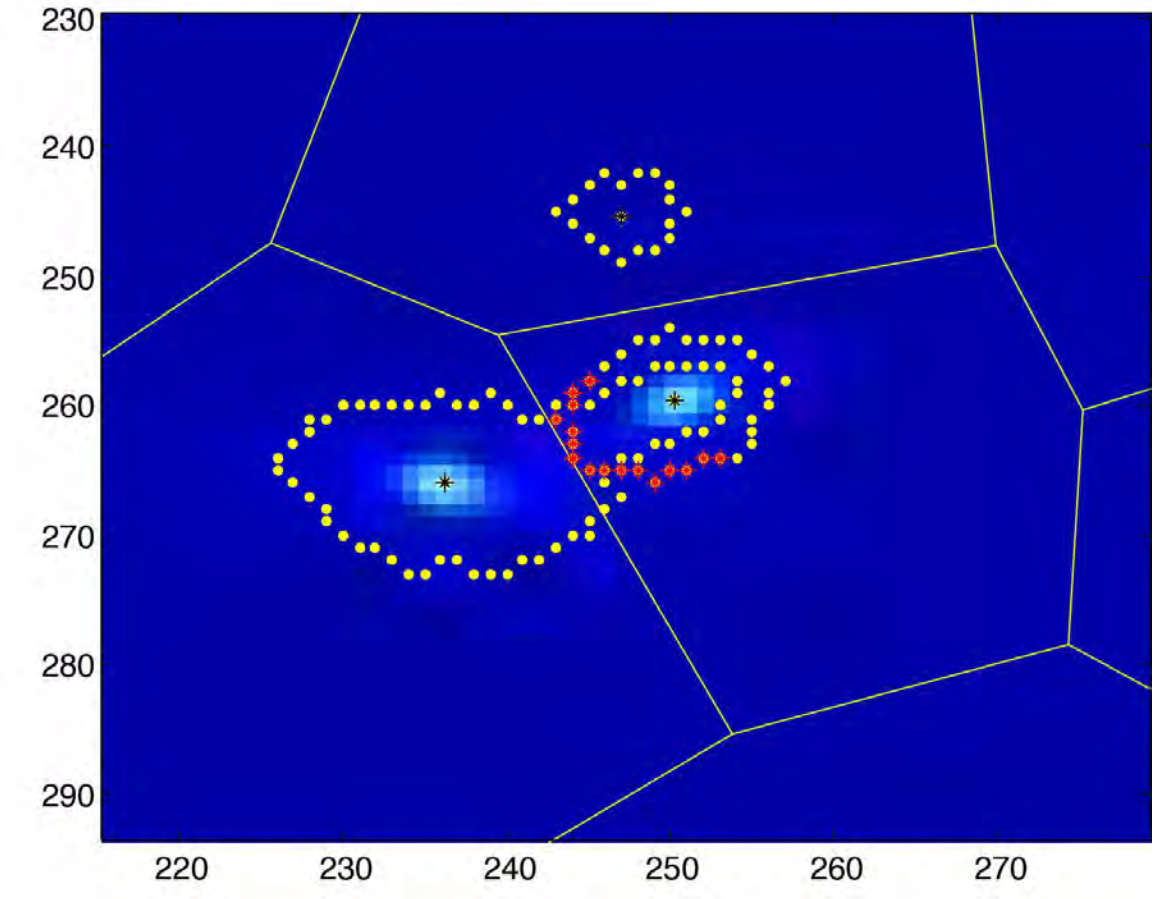

Figure 6.1: Frontier identified between two overlapped sources of an image with 88 artificial galaxies created using a Hubble distribution, Schecter luminosity function, and a signal-to-noise gain of 6. The red stars drawn on top of the yellow contours determine the selected frontier between both galaxies.

frontier among blended sources. For a further detailed study we can propose a combined Labelling - Active Contour algorithm. The goal is to dynamically adapt the active contour parameters λ_1, λ_2 and number of iterations based on the knowledge of the source labelling as isolated versus non-

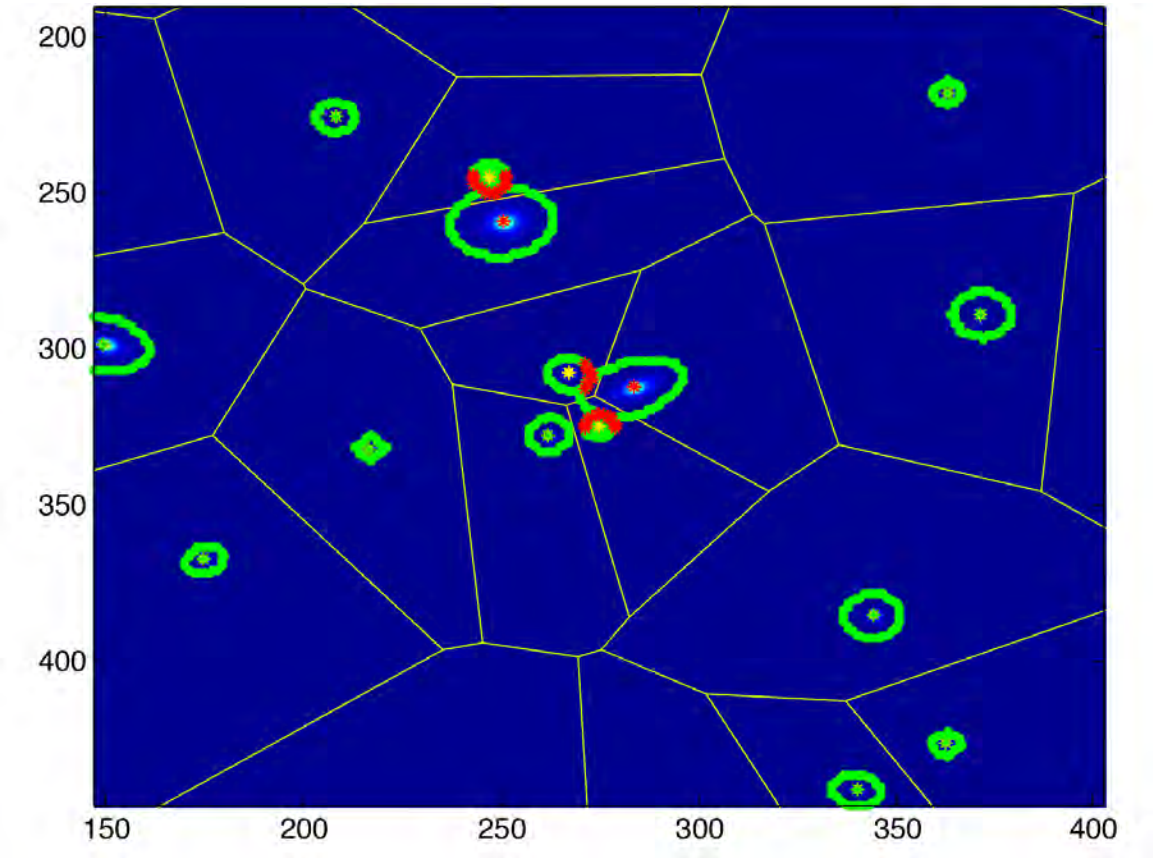

Figure 6.2: Frontier identified between several overlapped sources of an image with around 50 artificial galaxies created with a uniform distribution, Schecter luminosity function, and a signal-to-noise gain of 3. The red stars drawn on top of the green contours determine the selected frontier between both galaxies.

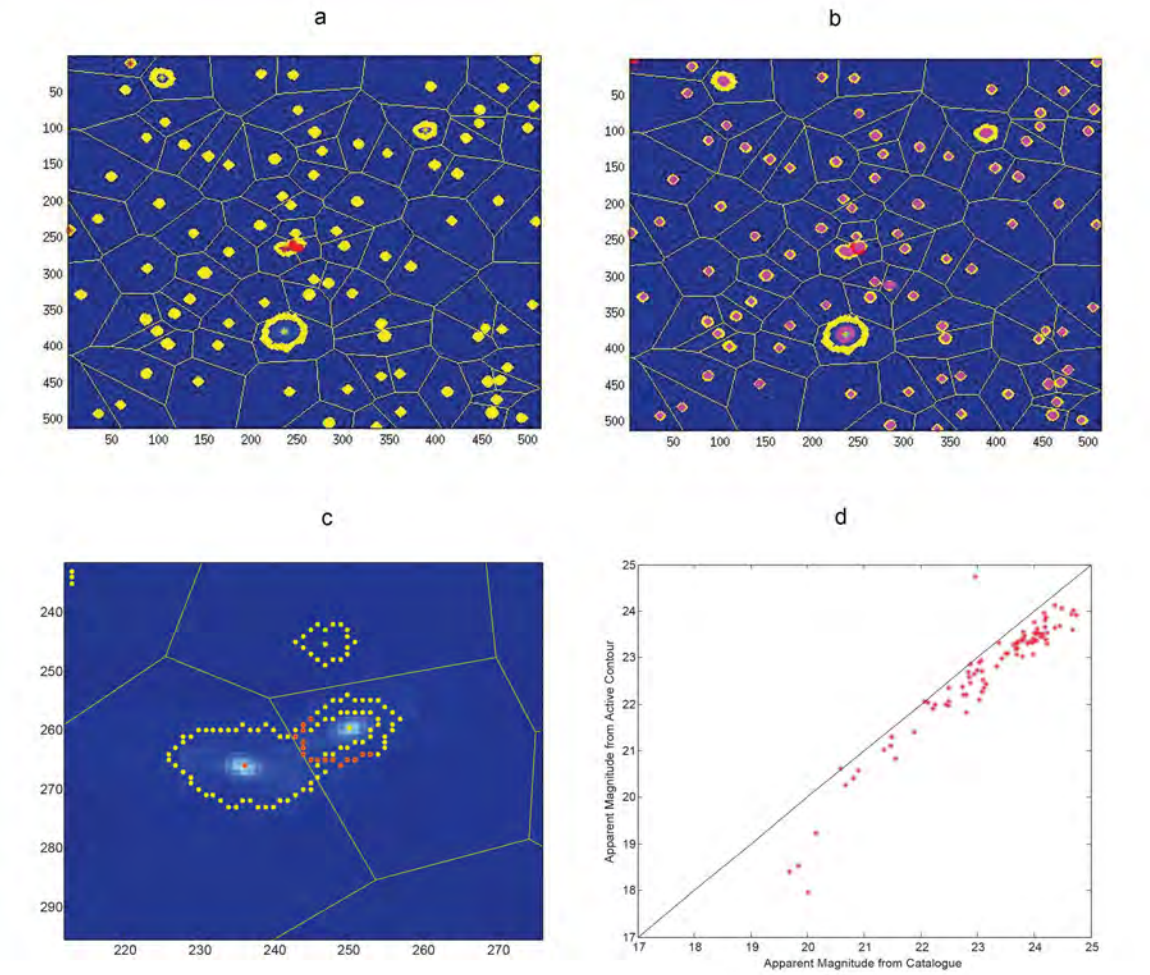

Figure 6.3: 88 IRAF artificial sources following Hubble distribution, with Poisson noise and $S/N = 6$. Top left panel. Voronoi tessellation and active contour (yellow). Top right panel. Voronoi tessellation showing the active contour (yellow) and the Kron ellipses (magenta). Bottom left panel. Zoom showing voronoi cells and active contours with blended sources in contiguous Voronoi cells. It also shows the preliminary promising results in determining the frontier between the blended sources. Bottom right panel. Apparent magnitude from SExtractor catalogue (x axis) and from active contour algorithm (y axis).

isolated.
The sequence of steps to implement would be the following one:

1 We obtain contours with an initial set of values for the parameters λ_1, λ_2 and number of iterations.

2 We compute the Voronoi tessellation

3 We label each source as isolated, contaminated or partially-contaminated, depending on the rules established in Section 3.8

4 Depending on the results from step [3] we accommodate the parameters λ_1, λ_2 and number of iterations.

Astro-Photo Cross Matching Pipeline The combination of astrometric and photometric information in order to improve the automatic knowledge of multi-band galaxies cross matching is definitively a good candidate for a future line of work. We already implemented the Bayesian inference for the astrometric cross matching problem, as described in [Budavári and Szalay, 2008]. The context of this work was the ASTRID project and this was the starting point of this thesis. Now, once the astrometric and photometric bayesian framework has been proven to be an effective addition to the galaxies cataloguing, the combination of both Bayes Factor will be an important optimization of the GCES in its aspect of the cross-matching. Figures 6.5 to 6.7present the results yielded for the astrometric cross matching solution implemented in the context of the project ASTRID. Figure 6.5 shows the outcome of labelling 586 sources following our rule-based system. Thus, red stars indicate contaminated sources, yellow stars correspond to partially contaminated sources and the isolated sources are indicated with green stars. Figure 6.5 shows what we named as Bayes Circles in a Voronoi tessellation done on celestial coordinates. These circles are the graphical represetation of group of sources candidates of being part the same galaxy. In other words, the astrometric Bayes Factor is higher than unity for the group of sources included in a Bayes circle. Figure 6.6 is a zoom of one of the Bayes circles of Figure 6.5. Here we can observe how the different grouping of sources determine different circles each of them with a corresponding Bayes Factor value. This correspondence with the astrometric Bayes Factor is presented graphically in Figure 6.7. This Bayes Factor tree allows us to assess various plausibilities path in the problem of astrometric cross matching.
Therefore, the addition of the photometric Bayes Factor in the framework presented is obviously feasible with minimum effort required. This line of work is intended to follow to the conclusion of this thesis.

Quality Flags Tree As conclusions, following the line of thought introduced at the beginning of this section and considering the overall picture of the results presented in this thesis, it seems logical to propose, as future line of work, the definition and use of three additional quality flags:

- Photo-match: this flag will measure the confidence in the cross-matching by adding the photometric knowledge..
- Confusion-label: this flag will indicate whether the source is labelled as isolated or as contaminated.
- Contour-accuracy: this flag is the result of the value of the confusion-label flag and the difference in magnitude values between the catalogue and the surface brightness algorithm of the GCES. Thus, if the confusion-flag indicates that the source is isolated and the GCES source magnitude value is smaller then the corresponding value of the catalogue, this means that the contour-accuracy is good to be reliable. Additional rules can be considered to cover other cases, however these rules will have to define a threshold value of the difference between the catalogue magnitude and the GCES magnitude.

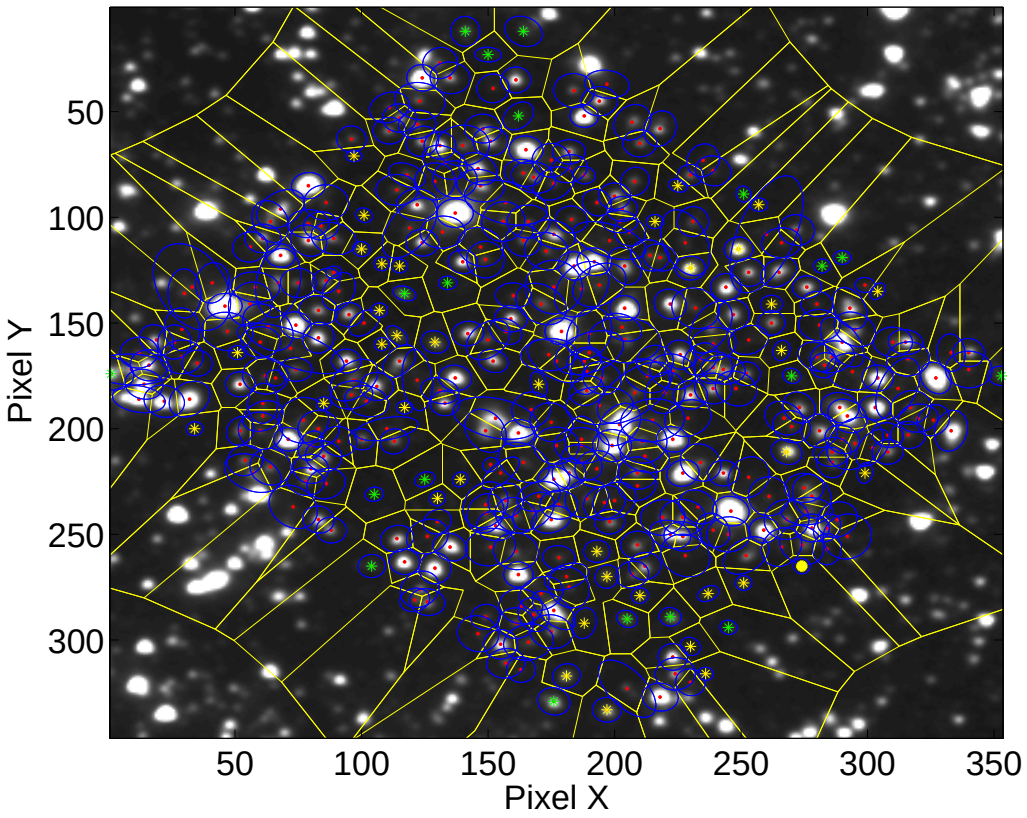

Figure 6.4: Labelling of 586 sources in SUBARU B,V,R,I and IRAC bands.

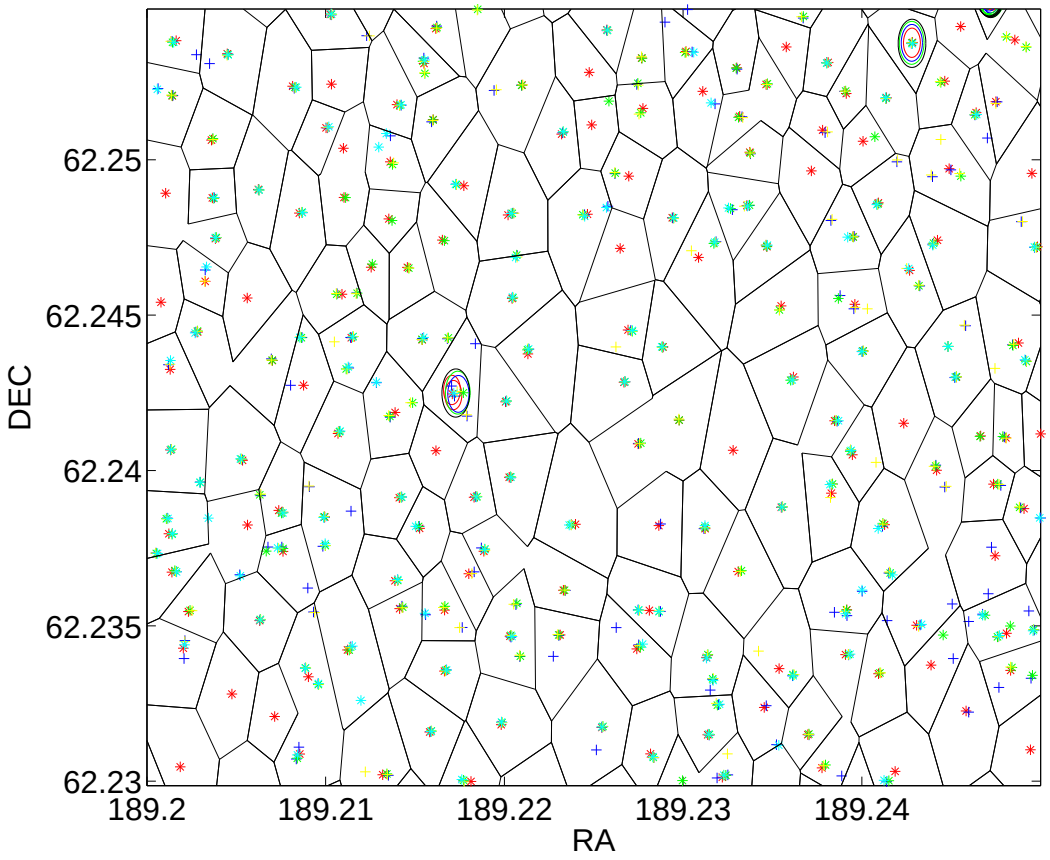

Figure 6.5: Astrometric Bayesian Cross Matching

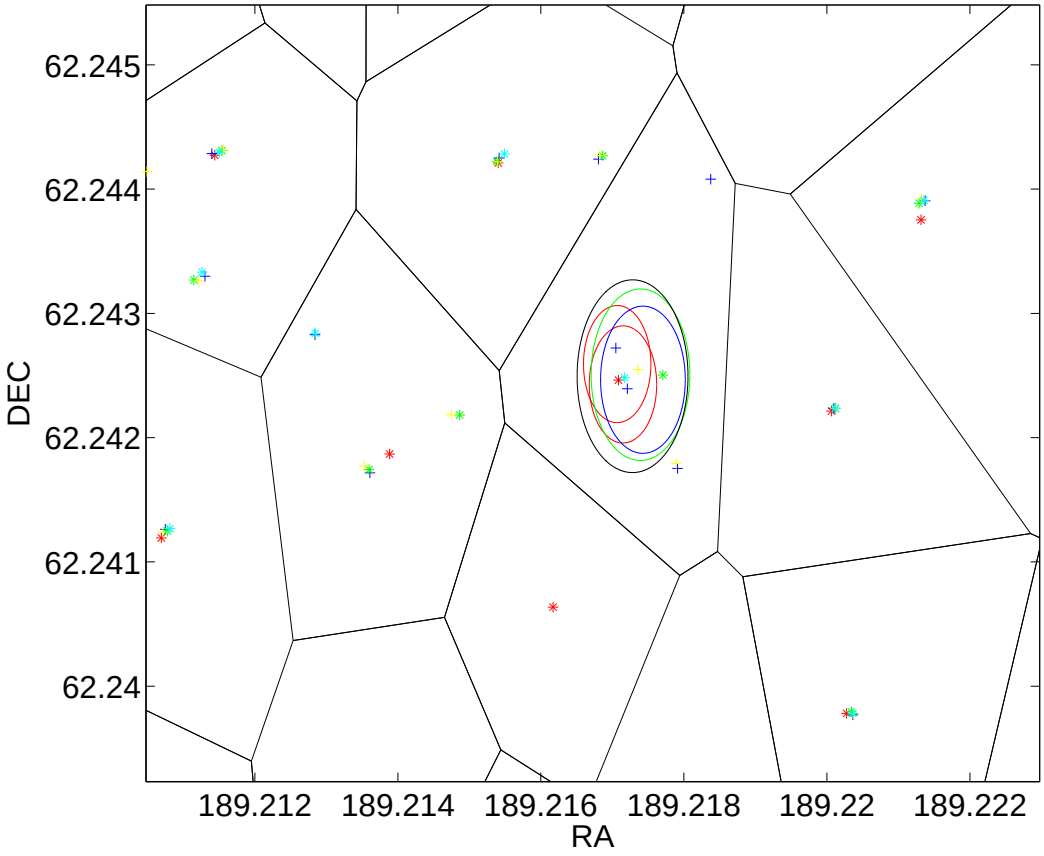

Figure 6.6: Astrometric Bayes Circle for 5 sources candidates of astrometric cross matching

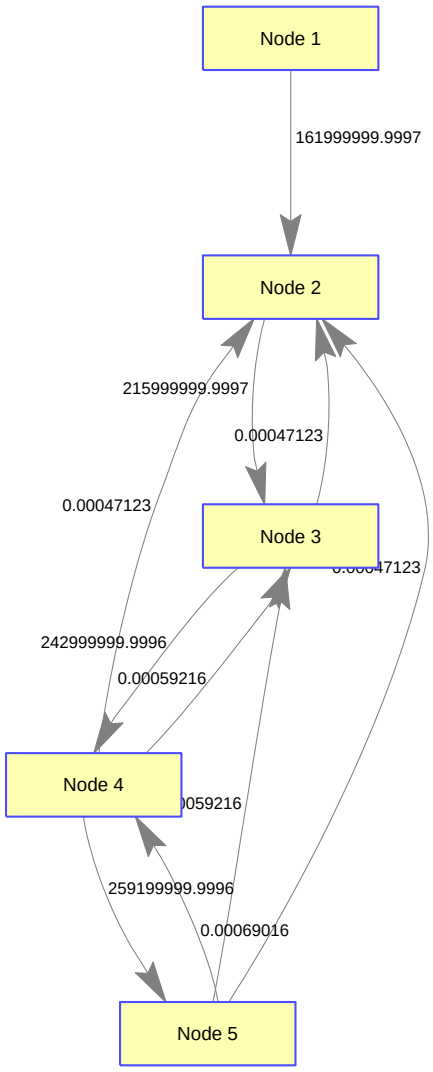

Figure 6.7: Astrometric Bayes Factor Tree

Today the cross-matching between catalogues is very extensive and diverse and therefore quality maps of the catalogues involved in the cross-matching are becoming more relevant. From that perspective, the notion of quality trees built up among different bands of different catalogues can provide a useful tool of information and further heuristic decision. For example, a quality flag can be low in one band of a survey but high in another band of the same or a different survey. If we build up quality flags trees for all the surveys, we can expand the concept of cross-matching by cross-assessing catalogues of different surveys but with equivalent quality flags, and this will obviously increase confidence in the scientific information extracted.

Bibliography

[Abazajian et al., 2003] Abazajian, K., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Anderson, S. F., Annis, J., Bahcall, N. A., Baldry, I. K., Bastian, S., Berlind, A., et al. (2003). The first data release of the sloan digital sky survey. *The Astronomical Journal*, 126(4):2081.

[Altman, 2014] Altman, Y. M. (2014). *Accelerating MATLAB Performance: 1001 tips to speed up MATLAB programs*. CRC Press.

[Arnouts and Ilbert, 2011] Arnouts, S. and Ilbert, O. (2011). Lephare: photometric analysis for redshift estimate. *Astrophysics Source Code Library*.

[Bailer-Jones, 2011] Bailer-Jones, C. A. (2011). Bayesian inference of stellar parameters and interstellar extinction using parallaxes and multiband photometry. *Monthly Notices of the Royal Astronomical Society*, 411(1):435–452.

[Ball and Brunner, 2010] Ball, N. M. and Brunner, R. J. (2010). Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, 19(07):1049–1106.

[Bechhofer, 2009] Bechhofer, S. (2009). Owl: Web ontology language. In *Encyclopedia of Database Systems*, pages 2008–2009. Springer.

[Bertin, 2001] Bertin, E. (2001). *Mining pixels: The extraction and classification of astronomical sources*. Springer.

[Bertin and Arnouts, 1996] Bertin, E. and Arnouts, S. (1996). Sextractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series*, 117(2):393–404.

[Beucher and Lantuejoul, 1979] Beucher, S. and Lantuejoul, C. (1979). Use of watersheds in contour detection. In *International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation, Rennes, France.*

[Beucher and Meyer, 1992] Beucher, S. and Meyer, F. (1992). The morphological approach to segmentation: the watershed transformation. *OPTICAL ENGINEERING-NEW YORK-MARCEL DEKKER INCORPORATED-*, 34:433–433.

[Boissonnat et al., 2006] Boissonnat, J.-D., Wormser, C., and Yvinec, M. (2006). Curved voronoi diagrams. In *Effective Computational Geometry for Curves and Surfaces*, pages 67–116. Springer.

[Bruzual and Charlot, 2003] Bruzual, G. and Charlot, S. (2003). Stellar population synthesis at the resolution of 2003. *Monthly Notices of the Royal Astronomical Society*, 344(4):1000–1028.

[Budavári et al., 2007] Budavári, T., Dobos, L., Szalay, A. S., Greene, G., Gray, J., and Rots, A. H. (2007). Footprint services for everyone. In *Astronomical Data Analysis Software and Systems XVI*, volume 376, page 559.

[Budavári and Szalay, 2008] Budavári, T. and Szalay, A. S. (2008). Probabilistic cross-identification of astronomical sources. *The Astrophysical Journal*, 679(1):301.

[Budavári et al., 2000] Budavári, T., Szalay, A. S., Connolly, A. J., Csabai, I., and Dickinson, M. (2000). Creating spectral templates from multicolor redshift surveys. *The Astronomical Journal*, 120(3):1588.

[Budding and Demircan, 2007] Budding, E. and Demircan, O. (2007). *Introduction to astronomical photometry*, volume 6. Cambridge University Press.

[Cambrésy et al., 2010] Cambrésy, L., Derriere, S., Padovani, P., Martinez-andrea, A. P., and Richard, A. (2010). Ontology of astronomical object types. *International Virtual Observatory Alliance. Available at: http://www. ivoa. net/Documents/Notes/AstrObjectOntology/20100117/NOTE-AstrObjectOntology-1.3-20100117. html. Accessed on January 2017*, 11(06):2011.

[Capak et al., 2007] Capak, P., Aussel, H., Ajiki, M., McCracken, H., Mobasher, B., Scoville, N., Shopbell, P., Taniguchi, Y., Thompson, D., Tribiano, S., et al. (2007). The first release cosmos optical and near-ir data and catalog. *The Astrophysical Journal Supplement Series*, 172(1):99.

[Caselles et al., 1997] Caselles, V., Kimmel, R., and Sapiro, G. (1997). Geodesic active contours. *International journal of computer vision*, 22(1):61–79.

[Chan and Vese, 1999] Chan, T. and Vese, L. (1999). An active contour model without edges. In *International Conference on Scale-Space Theories in Computer Vision*, pages 141–151. Springer.

[Chan and Vese, 2001] Chan, T. F. and Vese, L. A. (2001). A level set algorithm for minimizing the mumford-shah functional in image processing. In *Variational and Level Set*

Methods in Computer Vision, 2001. Proceedings. IEEE Workshop on, pages 161–168. IEEE.

[Chapin et al., 2004] Chapin, E. L., Hughes, D. H., and Aretxaga, I. (2004). A bayesian photometric redshift technique for mm-selected galaxies. In *Multiwavelength Cosmology*, pages 121–124. Springer.

[Coleman et al., 1980] Coleman, G., Wu, C.-C., and Weedman, D. (1980). Colors and magnitudes predicted for high redshift galaxies. *The Astrophysical Journal Supplement Series*, 43:393–416.

[Cristo et al., 2008] Cristo, A., Plaza, A., and Valencia, D. (2008). A novel thresholding method for automatically detecting stars in astronomical images. In *IEEE International Symposium on Signal Processing and Information Technology, 2008. ISSPIT 2008.*, pages 180–185. IEEE.

[Fadely et al., 2012] Fadely, R., Hogg, D. W., and Willman, B. (2012). Star-galaxy classification in multi-band optical imaging. *The Astrophysical Journal*, 760(1):15.

[Fatakdawala et al., 2010] Fatakdawala, H., Xu, J., Basavanhally, A., Bhanot, G., Ganesan, S., Feldman, M., Tomaszewski, J. E., and Madabhushi, A. (2010). Expectation–maximization-driven geodesic active contour with overlap resolution (emagacor): Application to lymphocyte segmentation on breast cancer histopathology. *IEEE Transactions on Biomedical Engineering*, 57(7):1676–1689.

[Feigenbaum and McCorduck, 1983] Feigenbaum, E. A. and McCorduck, P. (1983). *The fifth generation*. Addison-Wesley Pub.

[Fukugita et al., 1995] Fukugita, M., Shimasaku, K., and Ichikawa, T. (1995). Galaxy colors in various photometric band systems. *Publications of the Astronomical Society of the Pacific*, 107(716):945.

[Gaia Collaboration et al., 2017] Gaia Collaboration, Clementini, G., Eyer, L., Ripepi, V., Marconi, M., Muraveva, T., Garofalo, A., Sarro, L. M., Palmer, M., Luri, X., and et al. (2017). Gaia Data Release 1. Testing the parallaxes with local Cepheids and RR Lyrae stars. *ArXiv e-prints*.

[Gauch, 1999] Gauch, J. M. (1999). Image segmentation and analysis via multiscale gradient watershed hierarchies. *IEEE Transactions on Image Processing*, 8(1):69–79.

[Gorski et al., 2005] Gorski, K. M., Hivon, E., Banday, A., Wandelt, B. D., Hansen, F. K., Reinecke, M., and Bartelmann, M. (2005). Healpix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759.

[Gray et al., 2008] Gray, A., Gray, N., Hessman, F., and Preite Martinez, A. (2008). Vocabularies in the virtual observatory. *IVOA Proposed Recommendation http://www. ivoa. net/Documents/latest/Vocabularies. html (Accessed on January 2017)*.

[Gregory, 2010] Gregory, P. (2010). Bayesian logical data analysis for the physical sciences. *Bayesian Logical Data Analysis for the Physical Sciences, by Phil Gregory, Cambridge, UK: Cambridge University Press, 2010*.

[Hasanah et al., 2013] Hasanah, L., Iryanti, M., Ardhi, N. D., and Feranie, S. (2013). Devel-

opment of software for making contour plot using matlab to be used for teaching purpose. *Applied Physics Research*, 5(1):78.

[Heinis et al., 2009] Heinis, S., Budavári, T., and Szalay, A. S. (2009). Cross-identification performance from simulated detections: Galex and sdss. *The Astrophysical Journal*, 705(1):739.

[Hobson, 2010] Hobson, M. P. (2010). *Bayesian methods in cosmology*. Cambridge University Press.

[Howell, 2006] Howell, S. B. (2006). *Handbook of CCD astronomy*, volume 5. Cambridge University Press.

[Icke and van de Weygaert, 1987] Icke, V. and van de Weygaert, R. (1987). Fragmenting the universe. *Astronomy and Astrophysics*, 184:16–32.

[Ilbert et al., 2009] Ilbert, O., Capak, P., Salvato, M., Aussel, H., McCracken, H., Sanders, D., Scoville, N., Kartaltepe, J., Arnouts, S., Floc'h, L., et al. (2009). Cosmos photometric redshifts with 30-bands for 2-deg$^{\wedge}$ 2. *Astrophysical Journal*, 690(2):1236–1249.

[Jäger, 1990] Jäger, M. (1990). Hubble space telescope.

[Jeffreys, 1998] Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

[Kamath, 2009] Kamath, C. (2009). *Scientific data mining: a practical perspective*. SIAM.

[Kass et al., 1988] Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331.

[Kerekes et al., 2010] Kerekes, G., Budavári, T., and Csabai, I. (2010). Bayesian approach for matching multiple stellar observations. In *Journal of Physics: Conference Series*, volume 218, page 012012. IOP Publishing.

[Kiang, 1966] Kiang, T. (1966). Random fragmentation in two and three dimensions. *Zeitschrift fur Astrophysik*, 64:433.

[Krughoff et al., 2005] Krughoff, K., Connolly, A., Colberg, J., O'Mullane, W., and Williams, R. (2005). A source extraction web service with cross matching capability. In *Astronomical Data Analysis Software and Systems XIV*, volume 347, page 355.

[Laigle et al., 2016] Laigle, C., McCracken, H., Ilbert, O., Hsieh, B., Davidzon, I., Capak, P., Hasinger, G., Silverman, J., Pichon, C., Coupon, J., et al. (2016). The cosmos2015 catalog: Exploring the 1¡ z¡ 6 universe with half a million galaxies. *The Astrophysical Journal Supplement Series*, 224(2):24.

[Loredo, 1990] Loredo, T. J. (1990). From laplace to supernova sn 1987a: Bayesian inference in astrophysics. In *Maximum entropy and Bayesian methods*, pages 81–142. Springer.

[Lubow et al., 2011] Lubow, S., Budavári, T., and Cole, N. (2011). Cross matching sources in the hubble legacy archive (hla). In *Astronomical Data Analysis Software and Systems XX*, volume 442, page 97.

[Malik et al., 2003] Malik, T., Szalay, A. S., Budavari, T., and Thakar, A. R. (2003). Skyquery: A web service approach to federate databases. In *In Proc. CIDR*. Citeseer.

[Marquez et al., 2014] Marquez, M., Budavári, T., and Sarro, L. (2014). Improving cross-identification of galaxies using their photometry. *Astronomy & Astrophysics*, 563:A14.

[Marquez, 2012] Marquez, M. J. (2012). A new approach to the optimization of the extraction of astrometric and photometric information from multi-wavelength images in cosmological fields. In *Astrostatistics and Data Mining*, pages 181–189. Springer.

[Marquez and Sarro, 2013] Marquez, M. J. and Sarro, L. M. (2013). Data modelling and calibration using a two level hierarchical bayesian approach. *The International Journal of Soft Computing and Software Engineering (JSCSE)*, 3(3):98.

[Martin et al., 2005] Martin, D. C., Fanson, J., Schiminovich, D., Morrissey, P., Friedman, P. G., Barlow, T. A., Conrow, T., Grange, R., Jelinsky, P. N., Milliard, B., et al. (2005). The galaxy evolution explorer: A space ultraviolet survey mission. *The Astrophysical Journal Letters*, 619(1):L1.

[Matsuda and Shima, 1984] Matsuda, T. and Shima, E. (1984). Topology of supercluster-void structure. *Progress of theoretical physics*, 71(4):855–858.

[Matsumoto and Nishimura, 1998] Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30.

[McGuinness et al., 2004] McGuinness, D. L., Van Harmelen, F., et al. (2004). Owl web ontology language overview. *W3C recommendation*, 10(10):2004.

[McKeeman, 2008] McKeeman, B. (2008). Measuring matlab performance.

[Melott et al., 1983] Melott, A. L., Einasto, J., Saar, E., Suisalu, I., Klypin, A. A., and Shandarin, S. F. (1983). Cluster analysis of the nonlinear evolution of large-scale structure in an axion/gravitino/photino-dominated universe. *Physical Review Letters*, 51(10):935.

[Metchev et al., 2008] Metchev, S. A., Kirkpatrick, J. D., Berriman, G. B., and Looper, D. (2008). A cross-match of 2mass and sdss: newly found l and t dwarfs and an estimate of the space density of t dwarfs. *The Astrophysical Journal*, 676(2):1281.

[Mira, 1995] Mira, J. (1995). *Aspectos básicos de la Inteligencia Artificial*. SANZ Y TORRES.

[Mumford and Shah, 1989] Mumford, D. and Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685.

[Nieto-Santisteban et al., 2007] Nieto-Santisteban, M. A., Thakar, A. R., and Szalay, A. S. (2007). Cross-matching very large datasets. In *National Science and Technology Council (NSTC) NASA Conference*.

[Nishimura and Matsumoto, 2002] Nishimura, T. and Matsumoto, M. (2002). A c-program for mt19937, with initialization improved 2002/1/26.

[Okabe, 1992] Okabe, A. (1992). *Spatial tessellations*. Wiley Online Library.

[Okumura, 2002] Okumura, H. (2002). A straightforward c++ implementation of mt19937.

[Orallo et al., 2004] Orallo, J., Quintana, M., and Ramírez, C. (2004). *Introducción a la minería de datos*. Fuera de colección Out of series. Editorial Alhambra S. A. (SP).

[Osher and Fedkiw, 2006] Osher, S. and Fedkiw, R. (2006). *Level set methods and dynamic implicit surfaces*, volume 153. Springer Science & Business Media.

[Osher and Sethian, 1988] Osher, S. and Sethian, J. A. (1988). Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79(1):12–49.

[Pan et al., 2011] Pan, N., Feng, Z., and Wang, M. (2011). A robust improved chan-vese model based on gaussian regularizing level set. In *2011 4th International Congress on Image and Signal Processing*, volume 3, pages 1150–1154.

[Pineau et al., 2011] Pineau, F.-X., Boch, T., and Derriere, S. (2011). Efficient and scalable cross-matching of (very) large catalogs. In *Astronomical Data Analysis Software and Systems XX*, volume 442, page 85.

[Polletta et al., 2007] Polletta, M., Tajer, M., Maraschi, L., Trinchieri, G., Lonsdale, C., Chiappetti, L., Andreon, S., Pierre, M., Le Fevre, O., Zamorani, G., et al. (2007). Spectral energy distributions of hard x-ray selected active galactic nuclei in the xmm-newton medium deep survey. *The Astrophysical Journal*, 663(1):81.

[Press et al., 2007] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press.

[Ramella et al., 2001] Ramella, M., Boschin, W., Fadda, D., and Nonino, M. (2001). Finding galaxy clusters using voronoi tessellations. *Astronomy & Astrophysics*, 368(3):776–786.

[Romanishin, 2006] Romanishin, W. (2006). *An Introduction to Astronomical Photometry Using CCDs*. CreateSpace Independent Publishing Platform (8 Aug. 2014).

[Roseboom et al., 2009] Roseboom, I. G., Oliver, S., Parkinson, D., and Vaccari, M. (2009). A new approach to multiwavelength associations of astronomical sources. *Monthly Notices of the Royal Astronomical Society*, 400(2):1062–1074.

[Rumbaugh et al., 2004] Rumbaugh, J., Jacobson, I., and Booch, G. (2004). *Unified modeling language reference manual, the*. Pearson Higher Education.

[Schneider, 2014] Schneider, P. (2014). *Extragalactic astronomy and cosmology: an introduction*. Springer.

[Schombert, 2011] Schombert, J. (2011). Archangel: Galaxy photometry system. *Astrophysics Source Code Library*.

[Scott et al., 2005] Scott, L. R., Clark, T., and Bagheri, B. (2005). *Scientific Parallel Computing*. Princeton University Press.

[Sethian, 1996] Sethian, J. A. (1996). A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595.

[Shimasaku et al., 2001] Shimasaku, K., Fukugita, M., Doi, M., Hamabe, M., Ichikawa, T., Okamura, S., Sekiguchi, M., Yasuda, N., Brinkmann, J., Csabai, I., et al. (2001). Statistical properties of bright galaxies in the sloan digital sky survey photometric system. *The Astronomical Journal*, 122(3):1238.

[Shobbrook and Shobbrook, 1993] Shobbrook, R. M. and Shobbrook, R. R. (1993). *Astronomy Thesaurus*. Research School of Astronomy and Astrophysics, Mount Stromlo Observatory.

[Sivia and Skilling, 2006] Sivia, D. and Skilling, J. (2006). *Data analysis: a Bayesian tutorial*. OUP Oxford.

[Skrutskie et al., 2006] Skrutskie, M., Cutri, R., Stiening, R., Weinberg, M., Schneider, S., Carpenter, J., Beichman, C., Capps, R., Chester, T., Elias, J., et al. (2006). The two micron all sky survey (2mass). *The Astronomical Journal*, 131(2):1163.

[Solomon and Breckon, 2011] Solomon, C. and Breckon, T. (2011). *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons.

[Trotta, 2008] Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49(2):71–104.

[van de Weygaert, 1994] van de Weygaert, R. (1994). Fragmenting the universe. 3: The constructions and statistics of 3-d voronoi tessellations. *Astronomy and Astrophysics*, 283:361–406.

[Verma et al., 2011] Verma, O. P., Singhal, P., Garg, S., and Chauhan, D. S. (2011). Edge detection using adaptive thresholding and ant colony optimization. In *2011 World Congress on Information and Communication Technologies*, pages 313–318.

[Walcher et al., 2011] Walcher, J., Groves, B., Budavári, T., and Dale, D. (2011). Fitting the integrated spectral energy distributions of galaxies. *Astrophysics and Space Science*, 331(1):1–51.

[Wall and Jenkins, 2012] Wall, J. V. and Jenkins, C. R. (2012). *Practical statistics for astronomers*. Cambridge University Press.

[White, 1997] White, R. L. (1997). Object classification in astronomical images. In *Statistical Challenges in Modern Astronomy II*, pages 135–151. Springer.

[Wu and Leahy, 1993] Wu, Z. and Leahy, R. (1993). An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 15(11):1101–1113.

[Xia et al., 2009] Xia, L., Cohen, S., Malhotra, S., Rhoads, J., Grogin, N., Hathi, N. P., Windhorst, R. A., Pirzkal, N., and Xu, C. (2009). Improved photometric redshifts with surface luminosity priors. *The Astronomical Journal*, 138(1):95.

[Yoshioka and Ikeuchi, 1989] Yoshioka, S. and Ikeuchi, S. (1989). The large-scale structure of the universe and the division of space. *The Astrophysical Journal*, 341:16–25.

[Zhao et al., 1996] Zhao, H.-K., Chan, T., Merriman, B., and Osher, S. (1996). A variational level set approach to multiphase motion. *Journal of computational physics*, 127(1):179–195.

Vita