Characterization of the variability of Young Stellar sources in Star-forming regions across the Galaxy

Nicolás Medina Peña

Supervisor: Dra. Jura Borissova. (UV)

Internal reviewer: Dr. Amelia Bayo (UV)

External reviewer: Dr. Phil Lucas (University of Hertfordshire, UK)

External reviewer: Dr. Dante Minniti (UNAB, Chile)

Tesis para optar al grado de Doctor en Astrofísica

Instituto de Física y Astronomía

Facultad de Ciencias

Universidad de Valparaíso

Noviembre, 2022

Valparaíso, Chile.

Al lector curioso

Creo que las primeras palabras escritas y puestas deliberadamente en este documento deben ser para aquellas personas que no son científicas, que probablemente hojearan este manuscrito sin el fin de estudiarlo a profundidad. Para ustedes, y luego de dudarlo toda la vida, me es grato afirmar que una de las pocas formas verdaderas de comprender la realidad es a través de dedicar la vida a su contemplación.

This thesis is solely my own composition,
except where specifically indicated in the text.

Nicolás Medina Peña
Noviembre, 2022
Valparaíso, Chile.

Agradecimientos

Agradezco por sobretodo a la profesora Jura Borissova por siempre haber creído en mi. Al cuerpo académico de la Universidad de Valparaíso por sus productivas discusiones y sincera retroalimentación. También a los administrativos y auxiliares miembros de la UV que proporcionaron las condiciones para poder usar los espacios, los servidores, las salas, oficinas y todos los espacios comunes de nuestra universidad.

Agradezco a mi familia, amigos y amigas por dejarme guiar por esas luces brillantes que adornan el cielo de esa manera majestuosa e indiferente; las mismas que de una manera u otra influenciaron importantes decisiones, y todos esos días de luz y oscuridad que trajeron.

Especialmente, quiero agradecer a Nicol Flores, por todo su cariño y amor en el tormentoso periodo de concebir, escribir y editar un trabajo de este tipo. Gracias por tu compañía.

Finalmente pero no menos importante, agradezco a la beca de doctorado nacional CONICYT/ANID, folio 21181952 por financiar este proyecto de investigación. Como también al Instituto Milenio de Astrofísica (MAS) por dejarme participar en sus actividades y ejercer como observador del grupo APOGEE-2 survey, alojado en el observatorio Las Campanas.

Gracias a cada persona que cruzó palabras sinceras conmigo alguna vez.

Abstract

This thesis presents observational and statistical study of the irregular variable stars in four star-forming regions of our galaxy, namely: IC 2944/2948 (SFR1); G305 (SFR2); NGC 6357/NGC 6334 (SFR3), and the Lagoon nebula complex (SFR4). Based on the *VISTA Variables in the Vía Láctea* (VVV) ESO large survey we aim to unveil and characterize the Young Stellar Object (YSOs) population by analyzing the K_s-band variability. An automated process was developed to generate and analyze more than 24 million unevenly sampled time series over five years between 2010 and 2015, projected on $\sim$20 deg^2 area. We analyzed and tested several variability indices (features) and concluded that the total amplitude ΔK_s and the η index can be used as principal variability indicators for this type of variability. Our simulations also show that to select reliable variable star candidates with low content of contaminant sources, such as saturated objects or bogus the thresholds of the amplitudes $\Delta K_s > 0.6$ mag and η values < 0.95 (except for a variable source inside the SFR1 with $\eta = 3.2$) is necessary. Our final catalog contains 1096 high-amplitude irregular variable stars distributed along the bulge and the south-plane of the galaxy. The 524 of them present photometric measurements in J- and H-bands allowing us to construct the color-color diagram and quantify their status. More than 58% show intrinsic reddening, thus their variability can be connected to the early stages of the YSOs (stages I-II). The spatial location of classified variables correlates with star-formation features in both visible and infrared bands. Overdensities of the sample are distributed over HII regions, IRDCs, and their surroundings. We identified more than 3000 periodic stars inside the VVV time series as a byproduct.

Contents

CHAPTER 1

Introduction

For the people who spend their lives inside the cities, the observation of the night sky, our **Milky Way** (MW), and their overwhelming properties is not an easy task. We need to go through the light pollution that offers our houses and streets and escape to the isolated highs of the mountains to evade the residual illumination that reaches our pictures, `FITS` files, and, more important, still: our eyes.

Curiosity for those quiet and twinkling lights in the sky, which appear every night after sunset, drives the first steps into astronomy. The fact of understanding that every single point in the dark sky could be a sun or even a galaxy gives us the perspective of which is the place and scale that we populate in this reality. All the culture and adventures of humankind are bounded to the solar system, and still, our Sun is a low-mass object (I guess "low mass star" is better.) ($1,99 \times 10^{30}$ [Kg]) and represents almost 99,8% of the total mass of our planetary system.

In comparison to the large structures of the MW, such as massive stars, open and globular stellar clusters, the inter-stellar media, and the **star-forming regions** (SFRs), our humble Sun is just one more star inside the statistics of the galaxy. If we go around these thoughts quickly, it seems that humanity is insignificant, unable (still) to dominate other planets, even less to reach closer

stars. However, if we turn around the telescopes and modify a bit their optical components, we will see completely new tiny worlds, as small as the magnification of the microscope can offer. Probably if they can have consciousness and doubts about their reality, from their microscopic point of view, they would observe the large structures that humanity made. Using critical thinking, they will conclude something similar about their small nature. In general, we learn from the beginning of our immersion in astronomy that stars are eternal and unchanging, out of our human timescales. However, when we observe a large sample of stellar sources, the probability of observing infrequent events happens in every evolutionary stage in stars, from born to death.

An analogy for observing different time scales can be established if we consider that humans are studied for other short-lived organisms with which we share the earth: Butterflies. Depending on the specie, a butterfly's life varies from days to almost a year. Suppose somehow a butterfly dedicates their live to observe the human race. The insect could argue that it could gain valuable information by observing the behavior of several humans. In that case, they will observe children, older people, students or walking persons on the streets.

A butterfly cannot observe the entire life of a human being directly. However, with their observations, the butterfly can conclude that the humans are equal to other specie. We are born, raised, and die but in a longer life cycle. Of course, many unknowns will remain about the complete human life process. A follow-up for interesting humans needs to be performed to understand and predict their destiny with some degree of certainty. The importance of observing stars in short timescales is to constrain and characterize short processes in stars, discover new properties of already known systems, and map the structure of our galaxy.

1.1 Understanding the changes of brightness in the stars

From early in history, the observation of the sky leads to the recognition of several types of variable sources. The understanding of the *shape of their light*

curves made it possible to reveal signatures and properties about the nature of their magnitude changes in time. Variable stars awake the interest of the astronomical community since the late's 1500. In 1572, the famous and singular astronomer Tycho Brahe noticed the outburst of a bright supernova in the surrounding Cassiopeia constellation. He thought that he had contemplated the appearance of a new star, and in that way, it was reported.

Later, in 1596, David Fabricius observed another star's increasing and fading brightness inside the Cetus constellation. Fabricius also thought that he had discovered a new star as Tycho did, but after a while, he noticed with surprise that its brightness came back to normal after the star completely vanished. The same star discovered for Fabricius was first called *o* Ceti for Johann Bayer (1572-1625) and later called in 1662 "Historiola Mirae "(or *wonderful* in Latin) by Johannes Hevelius (1611-1687). Nevertheless, a third astronomer, Ismaél Bullialdus (1605-1694), noticed that Mira reaches their maximum brightness a month earlier every year, determining that the brightness change should be cyclic with a period of around 333 days. Today the reported period of *o* Ceti is 331.96 days.

This remarkable precedent established that an object in the sky could be interpreted, understood, and predicted, given their study and observation over time.

Four hundred years later, we already know that variability in stellar sources is not only observed in evolved stars. In the early stages of evolution, stars undergo rapid structural changes, and variations in their luminosity have been widely observed and reported. In order to deal with the categorization of every kind of variable source, it is necessary to define the main properties of variability.

Time-varying phenomena are arguably one of the most substantial and powerful sources of astrophysical information. Their study has led to many fundamental discoveries in modern astronomy in the last decades.

The ever-growing interest in time changing astronomical processes, as well as the tremendous development in astronomical instrumentation and automation during the past two decades, have been giving rise to several time-domain surveys of increasing scale. Wide-field shallow optical imaging surveys us-

ing small, dedicated telescope systems have been scanning the sky since the early 2000s with aims ranging from comprehensive stellar variability searches to exoplanet hunting, such as the Northern Sky Variability Survey (NSVS, (Hoffman et al., 2009)), Massive Astrophysical Compact Halo Objects Survey (MACHO, (Alcock et al., 1993)), the Optical Gravitational Lensing Experiment (OGLE, (Udalski, 2003)), the Palomar Transient Factory (PTF, (Law et al., 2009)), the wide-field optical imaging surveys: the Catalina Real-time Transient Survey (Drake et al., 2009); Pan-STARRS (Kaiser et al., 2002); and Gaia (Perryman, 2005).

In the near future, even more ambitious programs, such as the Large Synoptic Survey Telescope (LSST, Krabbendam & Sweeney (2010)), are planned to start monitoring the optical sky. While optical surveys are getting broader and deeper, the extension and the systematic exploration of the variable sky toward other wavelengths, such as the infrared, is also indispensable, not only for a more complete understanding of the observed phenomena, but also for overcoming the problem of interstellar extinction.

These synoptic surveys are expected to be the new powerhouses of modern astronomy by providing a high data flow for a wide range of science applications, from the study of transients to understanding the dynamics of the Milky Way galaxy.

1.1.1 Main types of stellar variability

Given the shape of their light curves, the general classification can be assigned for three main types: **regular** variability, **semi-regular** variability and **irregular** variability. First two basic categories emphasizes the presence of a periodic or at least quasi-periodic mechanism that leads the variation of brightness. Irregular sources display not celar signs of periodicity, but the variations of brightness could be prominent.

Nevertheless, stars could experiment with several causes that may vary their superficial brightness. Different variability classes have been discovered and gathered through the years and usually are represented in summaries called popularly “Variability trees“.

Image 1.1 provide one of the more popular of them, including stellar, extragalactic sources and asteroids. The most fundamental difference between their variability is to determine if the physical process related to the variability is provoked by the star or is caused by an external process, where the radiation from the source is modified when it reaches the observer. Variable star sources can be split into two main categories:

1. **Intrinsic variable sources**: The variability is inherent to the star, produced by a physical process associated with the observed source. The luminosity changes may drive changes in the properties of the system.

2. **Extrinsic variable sources**: The variability is external to the source. Apparent changes in brightness can be explained by modifying the quantity of light while it travels from the origins to the observer. This radiation may be altered given several factors, such as magnification, blocking, or scattering.

Starting from this point, it is possible to categorize every variable source that populate the skies.

1.2 The processes of star formation

Stars form from the gas and dust in molecular clouds via a series of complex processes that are currently only partly understood. Some fairly reliable empirical scaling laws exist, and a believable scenario exists for how an isolated low-mass star can form via turbulent fragmentation and disk accretion. Still, we lack a detailed understanding of how an entire molecular cloud turns into a cluster of stars. Most stars form in cluster environments, and stellar densities imply some interaction between **young stellar objects** (YSOs) during their formation. We must therefore analyze entire SFRs, at long enough wavelengths to not be limited by extinction and determine the physical properties of whole populations of YSOs.

Trends between the evolutionary state, mass, and spatial distribution of YSOs can help distinguish between star formation scenarios such as competi-

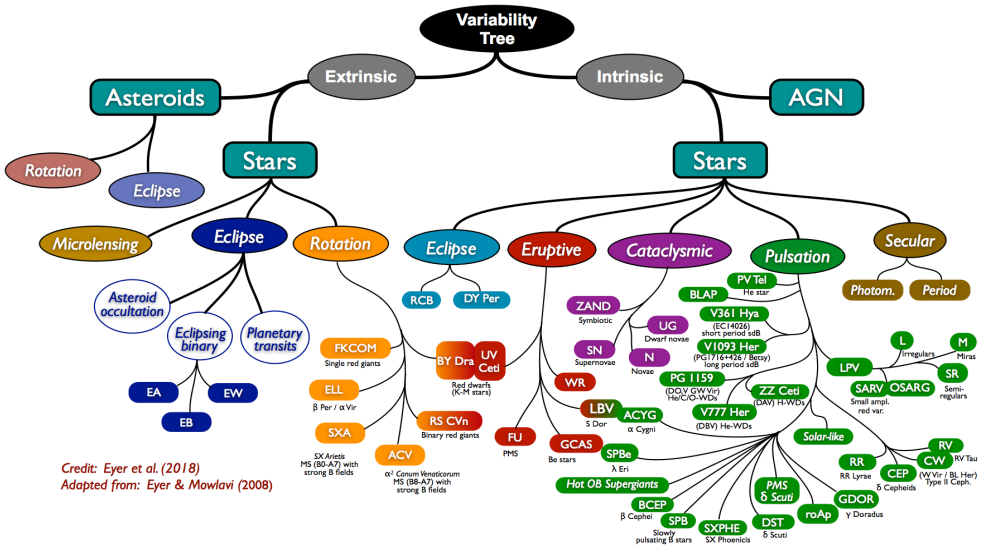


Figure 1.1: A proposed organization of variable objects and a visual summary of gathered information about the variability nature of astrophysical sources taken from Eyer & Mowlavi (2008), popularly called "the variability tree". Intrinsic and Extrinsic categories are the two main categories at the top of the plot.

tive accretion (e.g., Bonnell et al. (2004)) and sequential formation or triggering formation, which is troubling the complex and essential topic of star formation.

The formation of massive stars feeds energy back into the nearby interstellar medium, compressing, irradiating, and heating the natal molecular cloud. This feedback can have destructive or constructive effects, but it is unclear which dominates in a given cloud or overall in a galaxy. Although many examples exist of regions in our MW and other galaxies with suggestive spatial distributions of young and somewhat younger stars, determining the relevant physical conditions and timescales to make a strong argument for triggering is more complicated.

This issue must be clarified by assessing whether triggered star formation is happening in particular molecular clouds. It is necessary to know how the star formation proceeds and whether the stellar mass distribution that arises from the fragmentation of molecular cloud is the same for different molecular clouds or depends on the local conditions such as the entirely different stellar densities across the galaxy.

1.2.1 Basic properties of Star-forming regions

Inside the Galaxy are not only stars; there is a significant content of gas, dust, radiation, and matter in several states between the stars. The concentrations of interstellar material are observed within cloud-like structures, and this content is the material from which the stars are made. As a whole, it is commonly called the **interstellar medium** (ISM). The ISM is mainly composed of gas, containing a substantial fraction of atoms and molecules of hydrogen, a smaller fraction of Helium, and a small part of heavier elements such as oxygen, carbon, calcium, and iron, among several others.

Neutral Hydrogen HI clouds are cold ($T \sim 10$K), and their distribution can be detected using radio telescopes in the characteristic 21cm line, given their quantum properties. Denser regions inside those clouds allow the formation of molecular hydrogen H_2, carbon monoxide CO and complex molecules such as formaldehyde.

There are numerous cases of massive clusters forming at specific locations inside **giant molecular cloud** (GMC). Their spatial distribution suggests that the progenitor cloud may be filamentary, non-isotropic, and with some particular orientation. The elongated cloud structures may show an imprint on the clusters that form within them.

These filaments are cold molecular environments believed to be massive enough to drive the formation of eventually massive YSOs (i.e., $M > 8M_{\odot}$) with sufficient mass to produce detectable radio emissions associated with ionized hydrogen HII regions.

Similarly, young stars are formed deeply embedded inside dense filament in the GMCs. It is accepted that stars form by the gravitational collapse of cloud cores, and the main parameter for the understanding of the formation process and their theories is, in fact, the mass.

The density of these filaments is enough to display substantial light extinction even in IR, reaching a column density $\sim 10^{22}$ cm^{-2}, this is why they are called **infrared dark cloud** (IRDC). They can only be detected as absorption patches due to the background IR radiation from the milky way. It is necessary to use radio wavelength observations to go beyond these thick clouds.

Figure 1.2 displays an example of an IRDC localized in the bulge of the MW.

The collapse of massive molecular cores will produce signposts of massive star formation, such as the presence of Wolf-Rayet stars, which may have appreciably affected their natal environment. The shock front of the stellar wind emerges similarly as a bubble, expanding and compressing the ISM. This influence may trigger a new generation of stars with heavier elements in their compositions.

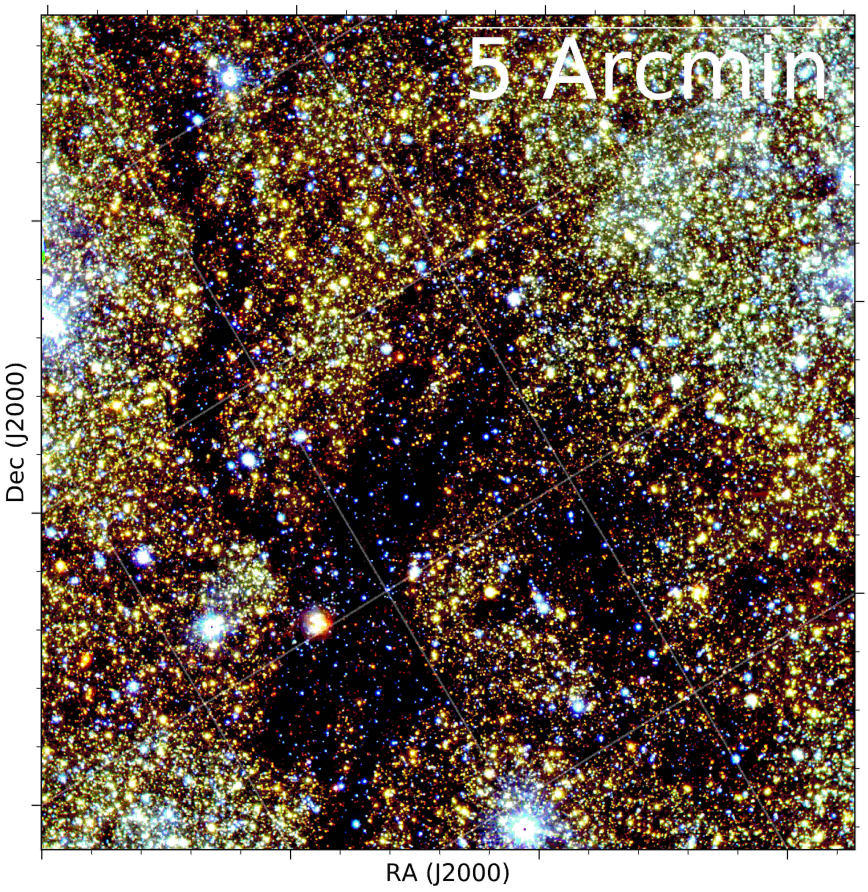


Figure 1.2: An example of an Infrared Dark Cloud (IRDC) distributed along the VVV tile b319. The image is centered at the coordinates (RA,Dec)=(266.840,−29.1986). The object inside the IRDC was called VVV CL148 (Borissova et al., 2014), a cluster candidate. Galactic north is up, Galactic east is to the left.

1.3 Young stellar objects

Various types of brightness changes have been noticed in YSOs. For instance, strong variability may be due to accretion events, while smaller-scale variability may be caused by geometrical effects, such as clearing of dust, spots, and rotation (Audard et al., 2014). FU Orionis stars (FUors; Herbig, 1966) are thought to be pre-main sequence (PMS) stars characterized a large increasing in brightness up to 4 magnitudes in optical wavelengths.

According to Audard et al. (2014), only about two dozen such sources were known in 2014. EX Lupi-type objects (EXors; Herbig, 1989) show similar brightness increases to FUors, but with shorter timescales. There are still problems in the interpretation of the variety of outburst intensities and timescales for those types of stars.

This variable brightness also can be observed in several stages of the classical model of PMS evolution, after the collapse of a molecular cloud has begun. Due to conservation of angular momentum, the infalling matter forms a protostellar disk, through which accretion on to the central source occurs (McKee & Ostriker, 2007; Li et al., 2014).

Numerous physical processes are involved in the evolution from a heavily obscured protostar until it reaches the main sequence. The physical origin of the variability could be a consequence of irregular intrinsic changes, such as episodic accretion (Audard et al., 2014), variable extinction induced by their circumstellar disks (e.g., Meyer et al. 1997; Cody et al. 2014; Rebull et al. 2014a), rotational modulation by cool and hot spots on the stellar surface, and the variations in accretion disk/envelope geometry.

Massive YSOs, with M$>$ 8M$_{\odot}$, likely undergo slightly different accretion and formation processes (Davies et al., 2011; Mottram et al., 2011; Frost et al., 2021). Observations of variability may provide some insight into these processes (Caratti o Garatti et al., 2017). Rice et al. (2015) determined that the amplitude of brightness variations is correlated with evolutionary class in YSOs. In other words, Class I YSOs are more variable than Class II, which are in turn more variable than Class III. Therefore YSOs at earlier evolutionary stages will be preferentially identified in variability-based searches.

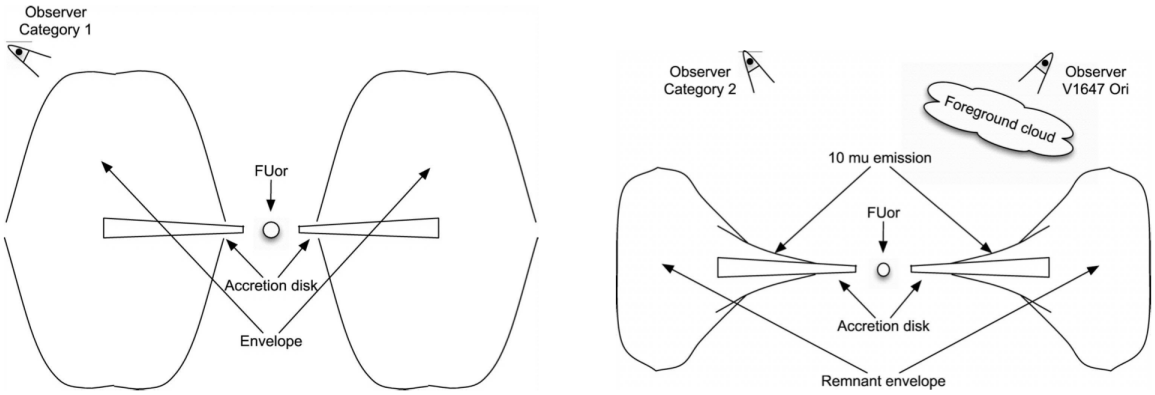


Figure 1.3: Sketch the two categories of FUors objects and geometry that can provoke variability, published in (Audard et al., 2014). The FUors categories are described in fully deails inside (Quanz et al., 2007). Mainly, The first category is related to embedded sources with silicate features in absorption and younger ages than FUors of the second category, which show silicate features in emission

However, the YSOs are often located inside or adjacent to dense molecular clouds (Evans et al., 2009) These clouds of material absorb and re-emit an important amount of the radiation from the young sources. Also, the disks and envelopes of the YSOs themselves are optically thick. Photometric surveys in the mid-infrared (MIR), such as Young Stellar Object VARiability (YSOVAR; Rebull et al., 2014b), or in the near-infrared (NIR) are very useful to minimize the influence of interstellar extinction.

The progenitors of YSOs are prestellar cores (starless cores, prestellar condensations). These are gravitationally bound, dense molecular cloud cores with typical stellar masses that may already be in a state of collapse, but have not formed a central protostellar object yet (Schmeja & Klessen, 2006).

Young Stellar Objects are stars in the first phase of theirs lives, before they enter the main sequence of the Hertzsprung-Russell diagram and are fed by stably hydrogen fusion. Thus, we can say that, YSOs are objects that are transforming from cloud cores to stars. They are formed by contraction (and fragmentation) of molecular clouds. Contraction can be started by a variety of factors, such as general density fluctuations in the interstellar medium, radiation pressure of nearby stars, or shock waves of supernova events that lead to

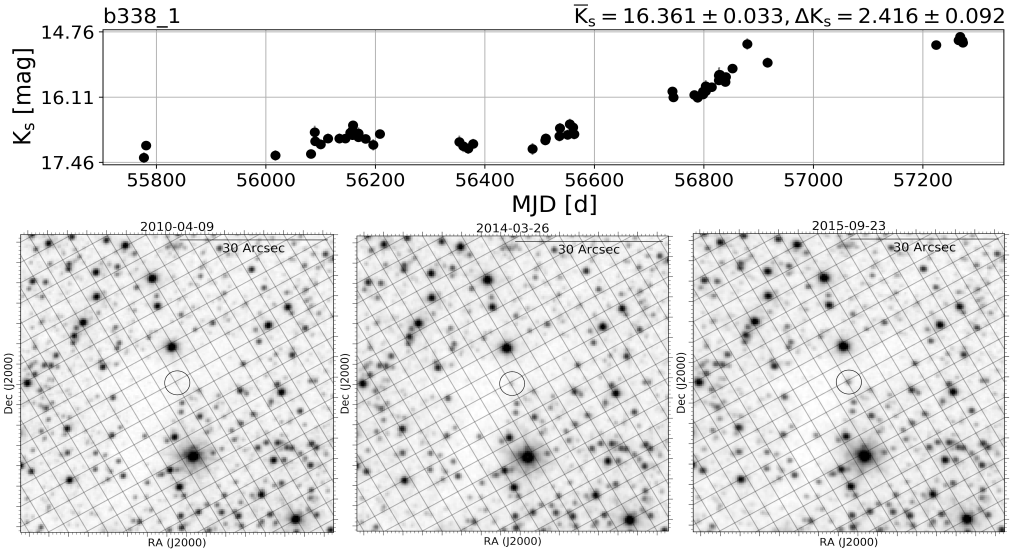
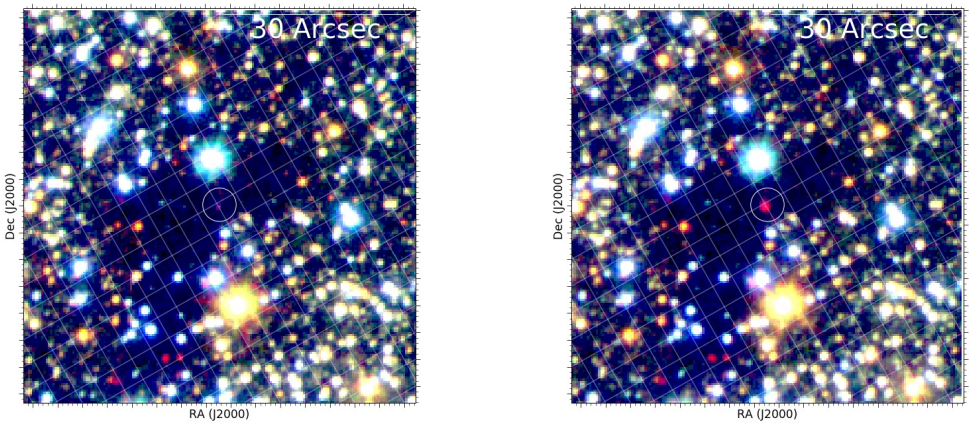


Figure 1.4: Visualization of the high-amplitude irregular variable star SFR4_235. Above plot represents the time-serie from VVV survey between 2010 and 2015. Bottom panels represent different epochs across the changes in brightness.



((a)) label 1 ((b)) label 2

Figure 1.5: Composed JHK$_s$ of variable star SFR4_235, same as displayed on figure1.4

local compressions.

The contraction of the molecular cloud is driven by gravity, the cloud actually collapses in free falling. The gravitational energy is released by radiation and in turn influences the collapse by its radiation pressure, which counteracts gravity. The dense center of the molecular cloud is the new protostar, which emits light due to the heat released by the gravitational collapse. Its core temperature, however, is still too low to maintain nuclear fusion.

In this protostar stage, the star is still growing by mass accretion from the surrounding molecular cloud, which lasts until either the entire cloud is incorporated or until the radiation pressure of the new star is powerful enough to blow off the remainders of the cloud. Using all this information, four classes of young stellar objects are proposed by (Lada, 1987), according to the properties of their spectral energy distributions (SEDs) (see, e.g., (Andre et al., 2000) for more details).

1.3.1 The Spectroscopic Energy distribution in YSOs

The radiation from the central star is absorbed by dust and re-emitted at long wavelengths. The different dust distributions among the YSO classes result in different spectral energy distributions. SEDs for classes I to III are characterized by the spectral index α. This slope α for a YSOs can be used to estimate its evolutionary stage in four different categories:

1. Class 0: According to Lada (1987), the Class 0 sources are deeply embedded protostars with a large sub-mm to bolometric luminosity ratio ($L_{smm}/L_{bol} > 0.005$). The Class 0 stage is the main accretion phase and lasts only a few 10^4 yr. As gas falls onto the surface of the protostar, it is heated in an accretion shock. This process converts the kinetic energy of the accreted gas into radiation. The photosphere is completely obscured by dust from the infalling envelope, but the luminosity can be measured from the reprocessed IR radiation. Class 0's are defined as having SEDs with roughly $T_{bol} < 70K$.

2. Class I: objects are relatively evolved protostars, which are surrounded by an accretion disk and a circumstellar envelope. Most of the accretion from the envelope is channelled through the disk. The accretion rate begins to decrease from the Class 0 stage. By the end of Class I, the protostar will have most of its final mass. The obscuration of the protostar by the envelope and disk is still significant ($A_V = 10 - 50$), although low enough that it can be detected at IR wavelengths, and occasionally in the optical if viewed near pole-on. SEDs that are rising ($\alpha > 0$) are often classified as Class I, although this boundary does not perfectly separate Classes I and II. The lifetime of Class I is roughly 0.1 Myr. The schematic model of this stage is shown in Figure 1.3.

3. Class II and III: Pre-main sequence stars correspond to the classical (CTTS) and weak line (WTTS) T Tauri stats, respectively. The classical T Tauri stage star, has an active disk, accompanied with formation of jets and Herbig-Haro objects, while the weak-line T Tauri star stage (WTTS) shows with a passive disk, the onset of nuclear fusion and fragmentation of the protoplanetary disk. Both classes II and III are characterized by a circumstellar disk (optically thick in Class II, optically thin in Class III) and the lack of a dense circumstellar envelope. Class II sources have noticeable accretion, but less than earlier stages, and too little to significantly increase the mass of the star. Without an envelope, Class II sources are obscured only by the molecular cloud, and edge on disks on rare occasions. As a result, they are often optically visible ($A_V = 0 - 5$), their SEDs exhibit excess emission ($\alpha > -2.5$). The classical T Tauri stars (CTTSs) are defined by strong enough Hα emission which indicate the accretion ($W_\lambda > 10A$) and they largely coincide with Class II objects. The average lifetime of Class II is roughly 3 Myr, although there is a large spread (1-10 Myr). Giant planet formation occurs during this stage.

4. Class III: YSOs are young stars without envelopes or optically thick disks. The star has negligible local dust and is obscured only by the molecular cloud. It often has had time to begin moving out of the cloud, further reducing its extinction. Giant planet formation has ceased, but the growth

of terrestrial planets can continue. The SED in this stage is dominated by the stellar photosphere, and shows no long wavelength excess emission, except perhaps a small amount from a newly formed debris disk. Weak line T Tauri stars (WTTSs) have weak Hα emission ($W_\alpha > 10A$) that suggests the absence of significant accretion. Some Class II sources are observed as WTTS, but Class III stars are always WTTS since they lack disks to supply accretion.

One of the most important properties of YSOs are that they are strongly correlated with star forming regions, nebulosity and dark clouds and they are variable. The variability can be connected with the star itself, which is highly variable for example during FU Ori-like phases and with the surrounding material, which is constantly changing during the formation process. For example, the dusty envelope around the star can casts erratic and temporally variable shadows on the surrounding molecular cloud. Often, these stars show irregular variability up to several magnitudes.

1.4 Time series in astronomy

Astronomy is an observational science. Unlike experimental sciences, we cannot interact with the object under study, change the conditions of the interaction or repeat an astrophysical event to obtain more data. The information we receive from the stars is generally limited by the light we can collect (although with gravitational waves this is not true). In general terms, we are able to understand the light in images, wavelengths, photon energy distribution, and time.

In the past, the astronomers were able to predict with a degree of certainty simply by assuming that those cycles would continue and could be measured. Astronomers of ancient times accumulated observations of solar, lunar, and planetary motion and eclipses for centuries and identified various cycles that had been repeated many times.

Early examples of astronomical time series include observing the position of stars and planets. Probably one of the first time series performed in modern

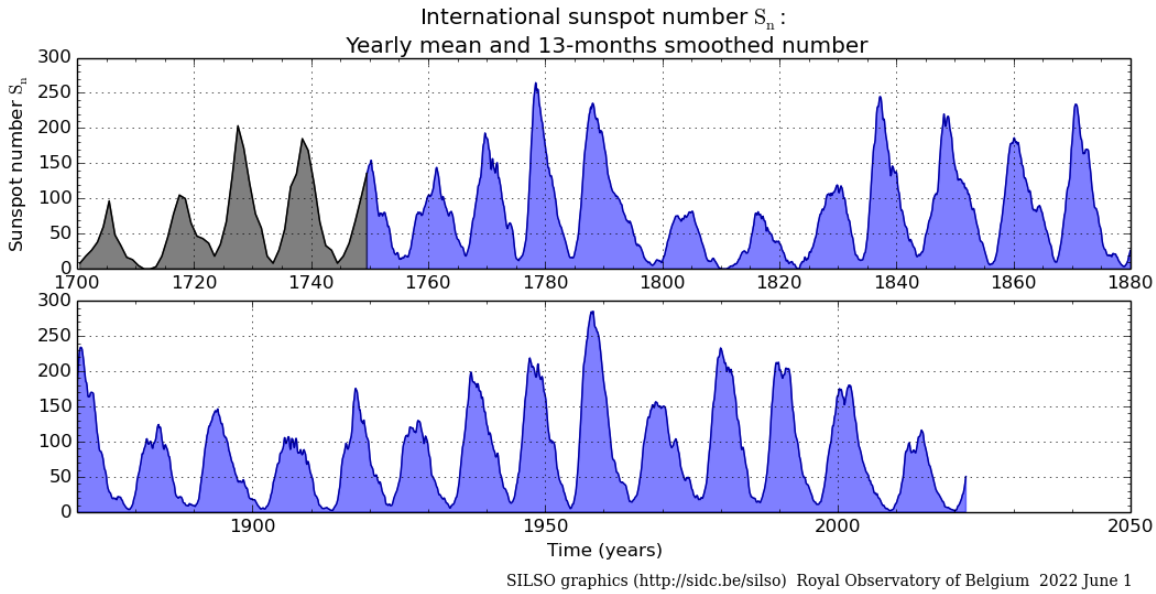


Figure 1.6: SILSO data/image, Royal Observatory of Belgium, Brussels

astronomy was Galileo's first Jupiter observations. He observed the intriguing appearance and disappearance of bright objects around Jupiter, and recorded the position of the moons of Jupiter over time.

A great example of an astronomical data series is the International Sunspot Number S_n (a.k.a. the Wolf number or Zürich number), which has been registered the number of sunspots per year since the advent of the telescope. In the year 1848, the astronomer Rudolf Wolf started to gather the observations from several sources in order to create the so-called the "Wolf number", which is the quantification of the number of individual and groups of sunspots on the surface of the Sun.

The original idea of creating a compilation of these values over time naturally produces a time series with a baseline of more than 300 years, which is still used today. Figure 1.6 displays the time series of the International Sunspot Number from the year 1700, and it is possible to see periodic features and trends between the spots number over the years. The plot was obtained from the SILSO database website[1] of the Royal Observatory of Belgium in Brussels.

[1] `https://www.sidc.be/silso/ssngraphics`

In the era of Big Data, astronomy is undergoing a major revolution due to the advances in technology. The astronomical surveys have evolved from taking observations from small and focused areas of the sky to wide-field surveys.

The Vera Rubin telescope was the project who alerted us to this massive availability of information in 2004, where they decide to install a 8.4m optical telescope which will be capable of imaging the whole sky twice a week. 15 TB per night would be the amount of information available and expecting for analysis, so the system to manage the data also should be massive and robust, allowing to produce alerts meanwhile some astronomical phenomena are taking place. (Jenness et al., 2016).

But not only the Vera Rubin telescope will be collecting data, will be the James Webb space telescope, the Giant Magellan Telescope (GMT), the EELT, etc. Before the COVID pandemic (2019), a Petabyte rate of data collection rate per night was estimated for 2022 only for astronomical purposes.

Basically, every detector around us is collecting tons of data from any source inside their range of sensibility; even our senses are also collecting data from the world at this very moment.

Time series are a well-known source of information to characterize and understand several branches of science. The key interest around time-series is to decode the information that is been release along the time of a certain parameter in study. Nevertheless, time domain in astronomy is different. The random and gaped nature of astronomical observations requires statistical attention to determine the reliability of the measurement. Also, data surveys need to adapt their set-ups to be able to receive and understand the radiation of certain astrophysical source in the sky.

1.4.1 Definition of time series

A time series can be describe as the sequence of consecutive observations of a certain parameter under measurement.

A formal definition of a time series is the application of a mathematical, statistical or any test in time-varying data, in order to quantify the variation itself of some particular parameter. This quantification is used to understand

and relate the behavior of the system, or at least get some idea of the variation in the system and their limits. The main goal of measure a time series is to obtain physical understanding in observation, what is the reason that makes the system variable, which are the differences and similitude between other variable systems. Therefore determine if it is possible to be able to predict the future behavior of the system using those data.

Considering that any aspect of reality that can be measured along the time has the potential to be a time series, the study of the physical processes, and the quantification of changes in the superficial brightness of sky sources can be measured and understood using time series.

If a time-dependent parameter $f(t)$ is measured at discrete intervals times t_i (with $i = 1, 2, 3, ..., N$), it is possible to construct a set of values $f(t_i)$ with uncertainty σ_{t_i} represented by:

$$\{t_i, f(t_i), \sigma_{t_i}\}. \tag{1.1}$$

This set of values in eq 1.1 is normally called a **time series.** The time t_i is taken such that the sequence $t_1, t_2, ..., t_N$ is strictly increasing. The time could me measured in arbitrary ranges and units. In astronomy/astrophysics is commonly to use the *Julian days* or some variation of this timescale. Table 1.1 represents the quick view of a typical time series taken from the photometry of a certain variable source in the sky.

1.4.2 Notation

In the literature, it is commonly found the use of several symbols and definitions for statistical analysis on time series. The use of time series is so widespread in the sciences that slightly depends on the field of your interest. All the time series have in common the natural temporal ordering in their measurements, hence it is necessary to discuss and adopt a nomenclature to refer to these quantities properly. For example, a time series could be simply defined as $f(t)$, displaying the basic idea of time-dependent function, which can collect measurements meanwhile time passes.

Modified Julian Day [days]	Magnitude [mag]	uncertainty [mag]
55250	14.546	0.161
55500	11.756	0.043
55750	12.241	0.046
56000	15.346	0.077
⋮	⋮	.⋮
57250	12.546	0.199

Table 1.1: Representations of the typical time series constructed from a time-varying astronomical source. The Modified Julian Day (MJD) is defined as MJD = JD - 2400000.5

Particularly in astronomy, the use of time series from celestial elements is referred to as a set of individual photometric magnitudes m_i, which were obtained at certain time t_i.

A time series is usually described as the vectorial quantity $X = \{x_n : n \in 1, 2, 3, ..., N\}$ or named simply $\{x_n\}$, where n refers to the total element number N inside the time series, normally called the index set. Sometimes, in the case of time series that are still added measurements to the set, the time series is called as $X = \{x_1, x_2, ..., \} = \{x_n\}_{n \geq 1}$.

1.4.3 Unevenly sampled time series

A time series is called regular if the measurements are taken or generated at regular intervals over time. If the time distance is not constant, or the measurement for the same parameter is taken with different cadences, the time series is **unevenly sampled**.

For ground-based telescopes, photometric time-series can only be sampled with non-regular intervals because of natural situations, such as the day/night cycle, or other inconvenient situations, such as telescopes committees.

This irregularly in time-domain data has been widely addressed in the past using spectral analysis to detect the presence of hidden harmonic components

that are caused by regular and semi-regular behavior brightness variations of periodic stars.

1.5 Periodic changes in variable stars

The analysis of time series produced by periodical astronomical phenomena is one of the most discussed in astronomy, given the critical information that the period represents in a sidereal source. Their study represents immense importance for astronomy. Pulsating variable stars are intrinsic variables, given that their variation in brightness is the manifestation of physical change within the star.

Main sequence stars such as our Sun are basically stable, exhibiting negligible changes in brightness (around 0.1% in the case of the Sun, in comparison with the Solar constant). These kinds of stars are in hydrostatic equilibrium: the gravitational force pulling the mass material to the center of the stars, meanwhile the radiation pressure generated by the nuclear reactions repels it, bringing about a balance between those forces, therefore to equilibrium.

In the case of pulsating variables stars, the periodic changes represent the expansion and contraction of surface layers around their core, leading to periodic increases and decreases in size. Perhaps the most valuable property of many types of pulsating variables is a direct relationship between the period of pulsation and their luminosity. This fact allows us to determine the distance between such stars and the distance to the galaxies that host them. Understanding the nature of their periodicity can provide us with information on the evolutionary state of the source or the physical system it represents.

The analysis of their time series provides vital information about the internal processes in stars. The different types of pulsating variables are distinguished by their periods of pulsation and the shapes of their **light curves** at different wavelengths. The analysis of their light curves can reveal the dominance of one period in the presence of several periods, making its nature even more complex.

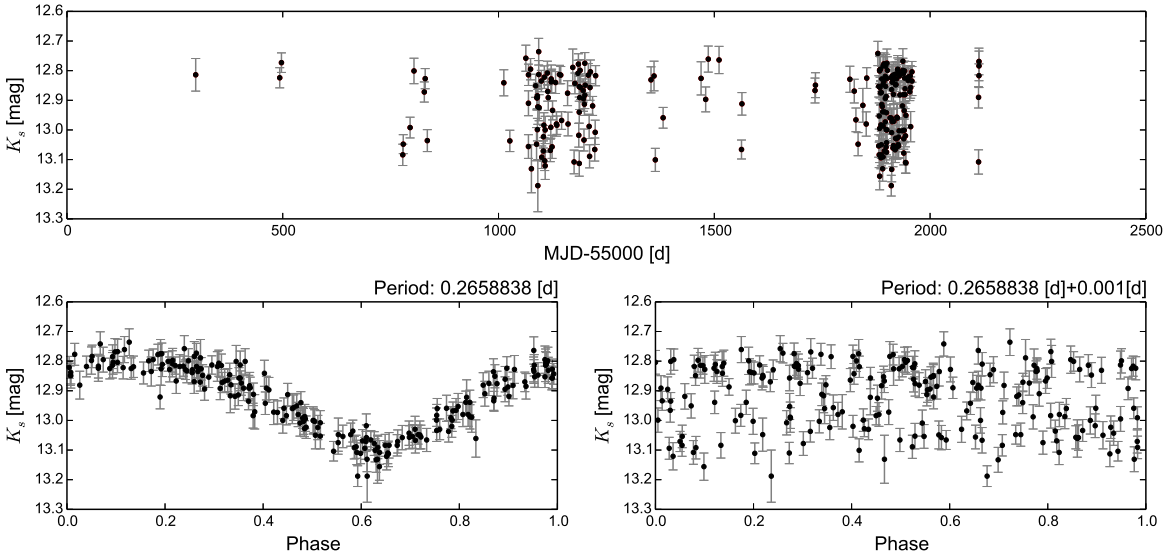


Figure 1.7: Visualization of the time series of a RRlyrae star. Above plot represents the time-series from VVV survey between 2010 and 2015. Bottom panels represent different epochs across the changes in brightness. Bottom-left plot: The time series is folded with a period $P = 0.2658838$ days, which is the fundamental period and therefore, maximizes the order of the time series in the phase space. Bottom-right plot: The time series was folded with an incorrect period $P + \delta P$, where $\delta P = 0.001$ days.

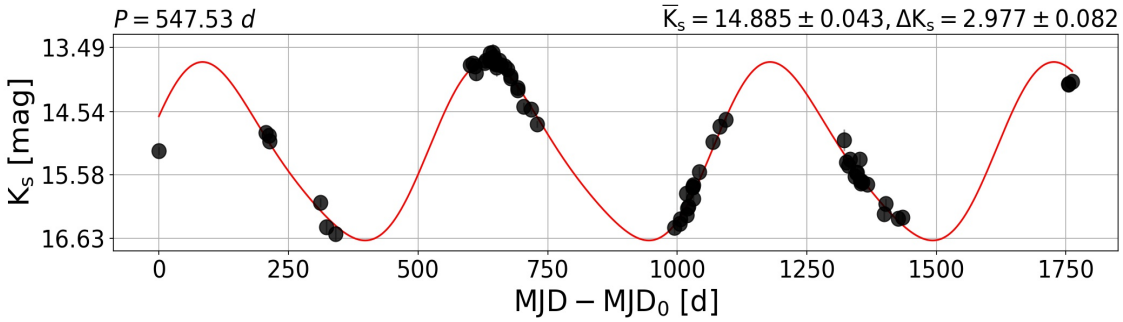


Figure 1.8: A typical time series of a long-period variable star. The regular variable cycle is visible along the photometric measurements for 1750 days. The red curve is a Fourier series modeled and fitted in the time series. Critical parameters such as the source period, the average K_s magnitude, and the total amplitude ΔK_s are displayed at the top.

1.6 Time series or light curves?

Historically the term **light curve** refers to the plot of the apparent magnitude of a specific variable star over a range of time.

Several authors and books use the term light curve as a synonym for the time series of a sidereal object. Nevertheless, the light curves are a particular case of the astronomical time series. More precisely, light curves are **phased time series**.

When a periodic star is monitored over time, it produces a regular time series, with fulfil the property $f(t) = f(t + P)$, where P is the main period of the source. A best way to represent a periodic star is using the **Epoch Folding technique**, which modifies the time component according to the following expression:

$$\phi_i(P) = \frac{t_i}{P} - \text{int}\left(\frac{t_i}{P}\right), \tag{1.2}$$

where $\phi_i(P)$ is the phase at the times t_i and int(t_i/P) is the integer part of (t_i/P).

This transformation is helpful given the gapped nature of astronomical time series. In general, it is enough to obtain all the flux measurements as possible (i.e., a randomly spaced time series), sort them to reconstruct the vari-

ability cycle, and describe how the magnitude changes in a single cycle. If P is the real period of the time series (or an integer multiple of it), the result of using equation 1.2 will be an ordered light curve in the phase space $\{\phi_i, m_i\}$.

Inside this phase space, the shape of the light curve can reveal information about the pulsation process of a periodic source. However, a small error δP will produce a significant disorder in the phase space, and the time series will look like noise. The effect of a small error in the period can be seen in figure 1.7.

Given the massive amount of data continuously taken by different surveys, in practice, a grid of possible periods P_i is evaluated inside equation 1.7 to evaluate the folded light curve with a variety of different techniques.

1.7 Motivation and general objectives

Shortly, space- and ground-based telescopes will have the potential to gather untold amounts of information and data sets. It will be possible to identify different properties with a certain degree of certainty, allowing the scientific community to make predictions from the data in real time. The hypothesis that emerges from here is that it will be possible to discriminate which are the main characteristics of certain types of astronomical objects.

Time domain surveys are full of surprises and are capable of understanding and potentially predicting the nature of millions of millions of stars following only their time series. The main tools proposed and applied in this thesis were gathered to extract information using several techniques of time-domain metrics. Therefore the primary motivation in analyzing time-domain surveys is to develop and create frameworks for the analysis of astronomical time series, to separate noisy and sparse time series data from accurate variable sources. Also, to gain experience managing and quantifying the computational resources needed for this task.

1.8 Specific objectives

The main focus of this study is to characterize the near infrared irregular variability across several star-formation regions in our Galaxy.

More specifically our goals are:

1. To improve the statistics of the YSOs, adding to the existing optical catalogs the near infrared variability.
2. To trace the star formation regions using newly discovered objects.
3. To connect the infrared variability with possible underlying physical processes.

1.9 Thesis structure

The thesis structure is:

1. The data used in this manuscript.
2. The targets under study.
3. The principal methods of analysis and ways to extract information.
4. Results
5. Summary

CHAPTER 2

Data

Infrared time-series photometry had long been constrained mostly to follow-up observations of known objects, but in recent years near-infrared imagers have been converging to optical CCDs in terms of resolution and performance, hence wide-field surveys became feasible with some NIR instruments, such as the Wide-Field near-IR CAMera (WFCAM, (Casali et al., 2007)) on the 3.8m United Kingdom Infrared Telescope (UKIRT), and the VISTA InfraRed CAMera (VIRCAM, (Dalton et al., 2006)) on the 4.1m Visible and Infrared Survey Telescope for Astronomy (VISTA), both hosting a variety of deep Galactic and extragalactic surveys. Nevertheless there are only a handful of wide-field NIR time-domain surveys that have more than a handful of observational epochs, with only the VISTA Variables in the Vía Láctea survey (VVV, (Minniti et al., 2010)) being comparable to the optical ones in terms of both area and time-domain coverage.

In this section, we will focus on the photometric data used in our work, namely the K_s band time series of individual stars. This include the process of measuring the fluxes for individual epochs and the methodology to extract several million time-series of every considered sideral source. the main source of photometric information is the VVV survey.

Along the years of development, the procedures described here were automatized and summarized in the flux diagram of Figure 2.5, their steps are also described, and their role in the analysis is discussed.

2.1 VVV survey

The VISTA Variables in the Vía Láctea survey (VVV; Minniti et al. 2010; Saito et al. 2012) is an ESO Infrared public survey that uses the 4-meter VISTA (Visible and Infrared Survey Telescope for Astronomy) telescope located in the Cerro Paranal observatory (Chile) for mapping 562 deg^2 in the Galactic bulge and the southern disk in five near-infrared broad bands: Z (0.87μm), Y (1.02μm), J (1.25μm), H (1.64μm) and K_s (2.14μm).

It has been designed to catalog $\sim 10^9$ sources, where a significant part of those are expected to be variable stars. It is comparable to the optical ones both in the area and time-domain coverage (e.g., Arnaboldi et al. (2007, 2012)).

The telescope has a near-infrared camera called VIRCAM that consists of an array of 16 detectors with 2048x2048 pixels and an average pixel scale of ≈ 0.34 [arcsecs/px]. The VVV survey also has a time coverage spanning over five years in the K_s band and produced a catalog of approx. 10^9 point sources (Minniti et al., 2010; Saito et al., 2010; Catelan et al., 2011) between 2010 and middle 2015.

The 16 detectors are spatially separated from each other, and a single exposure is called "pawprint", covering an area of 0.599 deg^2. To fill the gaps and obtain a contiguous image, six shifted pawprints are combined into a "tile" covering 1.5 x 1.1 deg, which in the case of VVV, are aligned along with Galactic l and b respectively. The total exposure time of a single tile is 80 seconds.

The observation was carried out between 2010 and the first semester of 2015. To cover the VVV survey area, the disk field is then divided into 152 tiles, while the bulge contains 396 tiles. The field of view of the survey and their tiles is shown in Figure 2.1

Some focused studies have been carried out in the southern disc region and the heavily crowded and reddened fields of the Galactic bulge, in

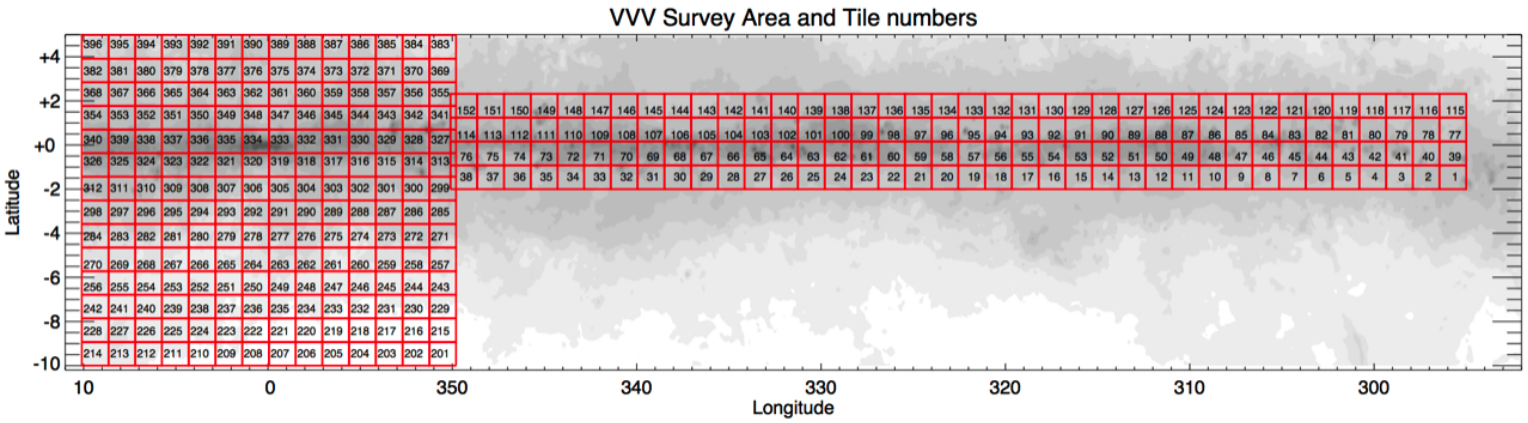


Figure 2.1: The VVV field of view and numbers of observed epochs taken of (Minniti et al., 2010)

search of variable stars using VVV time series, for example: Contreras Peña et al. (2017b), cataloging high amplitude variable stars, with emphasis on YSOs; Borissova et al. (2016) searching for YSOs around young stellar clusters; Dékány et al. (2015) searching Classical Cepheids in the bulge; Elorrieta et al. (2016); Gran et al. (2016); Minniti et al. (2017) focusing on the RR Lyrae stars. VIRAC (Smith et al., 2018), a near-infrared proper motion and parallax catalogue. VIVA (Ferreira Lopes et al., 2020) compiling the massive variability information for the heavily crowded and reddened regions of the Galactic Plane and Bulge; (Guo et al., 2020) analyzing the Short-and long-term near-infrared spectroscopic variability of eruptive protostars from VVV; among others.

All these sources are used to map the structure of the optically obscured Galactic disk and bulge by using distance indicators such as the red-clump giants and pulsating variable stars (RR Lyrae stars, classical Cepheids, anomalous Cepheids and Miras, and semi-regular variables), as well as to provide a census of Young Stellar Objects (YSOs) across the southern Galactic plane.

On the other hand, the VVV multi-epoch observations produced a massive amount of information, a dataset of challenging size. It is indispensable to develop tools, processes, and techniques that lead to performing sophisticated analysis in an automated way in order to exploit this unique dataset efficiently.

The Cambridge Astronomical Survey Unit[1] (CASU) combines and analyzes the data, creating the catalogues that provide parameters such as positions and fluxes from different found sources. All the images in J, H, and K_s bands were retrieved from the VISTA Science Archive[2] (VSA) database (Cross et al., 2012) provided that the image has been created without problems (i.e. flag "deprecated"=0) from pawprints.

2.1.1 PSF Photometry

The Point Spread Function (PSF) photometry of the data was fitted and performed using the software called DOPHOT (Schechter et al., 1993; Alonso-García et al., 2012). DOPHOT software expected mainly two values to perform the photometry properly: The atmospheric seeing at the observed night; and the estimated sky value. In both cases, DOPHOT documentation says that a value of 10% precision is good enough.

The astronomical seeing is a significant variable quantity that depends entirely on the weather conditions, and it is reported in the header of every VVV fits file. Images and photometric measurements estimated with seeing larger than four arcsecs were usually rejected; or contain a significant photometric uncertainty in their instrumental measurements. On the other hand, the sky value is determined for every individual image if we consider all the pixels values in the image lower than the average pixel values and determine the mode from the histogram, using the section 4.4.1.

There are two criteria to discriminate if the PSF photometry is accepted as valid given the output parameters: The "objtype" classification flag source, and the "Chi" or "χ" parameter. The First one refers of the DOPHOT classification of the extracted source, this classification is divided into nine categories, in which we consider only four: Sources with objtype= 1 (i.e., a perfect star), objtype= 2 (source shaped as a galaxy), objtype= 3 (close source) and objtype= 7 (faint source).

The second parameter represents the DAOPHOT (Stetson, 1987) parameter χ,

[1] http://casu.ast.cam.ac.uk/

[2] http://horus.roe.ac.uk/vsa/

which serves to quantify the photometric uncertainty of the source. This parameter is defined as the ratio of the actual deviation of the PSF fitted to the source to the expected deviation from the PSF calculated from the sky value and the readout noise of the image. For sources in which the PSF are well approximated, $\chi \approx 1$. If the source in the image is saturated or in the case of non stellar objects (among other issues), the χ parameter will be larger than unity.

If a extracted source is classified as objtype$= 1$, will be selected without counting their Chi value. In the case that objtype$= [2, 3, 7]$, sources with $\chi > 3$ are rejected because they have a very high photometric uncertainty associated (see e.g., (Bolte, 1987)). The selection scheme can be visualized on figure 2.2.

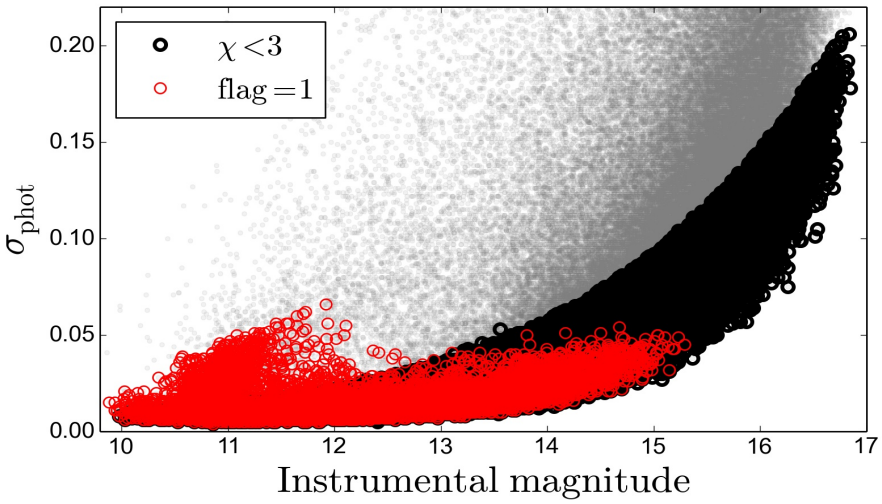


Figure 2.2: Example of the PSF photometric uncertainty σ_{phot} as a function of the instrumental DOPHOT magnitudes. On the plot we display the sources with $\chi < 3$ (open black circles), sources with $\chi > 3$ (transparent gray dots) and the special case with objtype flag$= 1$ (red open dots).

The output of the PSF process is an individual catalog with the physical coordinates (X,Y) (i.e., pixel coordinates) where the centroid of respective sources are estimated; the instrumental magnitude provided by the PSF photometry, and their associated magnitude uncertainty. The Right Ascension (RA) and Declination (Dec) coordinates for every sidereal source were determined using the WCSCTRAN routine from IRAF software, which performs a transformation

from the physical coordinates (X,Y) to the World Coordinate System (WCS). RA and Dec coordinates were estimated in degrees and with seven decimal places to preserve the high-quality astrometry provided by the VISTA telescope.

The PSF fitting and calibration was a process almost wholly designed for the analysis of VVV K_s-band, given their multiples epochs.

2.1.2 Calibration in the VISTA system

The calibration of our PSF photometry with the VISTA system was done using the aperture photometric catalogs made by CASU in the regions of interest. For this issue we are selected a group of standard stars in the FoV to compare the PSF and Aperture photometry.

To calculate the magnitudes in the VISTA system, we retrieved all the available aperture photometric catalogs of the studied regions between 2010 and 2015. Stars with CASU morphological classification flag "-1" (stellar sources) and "-2" (borderline stellar) were selected. Then we perform the cross-matching in all the catalogs using the `STILTS` package (Taylor, 2006) under the following conditions: 1.) The sources are within 0.4 arcsec, 2.) The sources appear on at least ten observations. 3.) The variation on the amplitude ΔK_s must be under 0.15 [mag], which can also be measured by selecting sources with a photometric standard deviation less than 0.03 magnitudes.

Typically, such constrained "standard stellar catalog" contains between 2000 and 10000 sources per tile. This standard star catalog is used to calibrate each of the PSF-generated catalogs from an available epoch within the tile (see figure 2.3).

Then, the VISTA K_s magnitudes were correlated with our PSF photometry using a least-squares linear regression, performed, in general, between a range of magnitudes $11 \leq K_s \leq 16$ for each epoch in this band in order to obtain the transformation coefficients. The calibrated magnitude m_{cal} is obtained from the linear fit:

$$m_{cal} = m_{\texttt{DOPHOT}} S + b \tag{2.1}$$

where S is the slope of the linear regression between the CASU and DOPHOT photometry, and b the intercept.

To avoid the outliers in the photometric measurements, we implemented the sigma clipping technique of one iteration, discussed on section 4.4.3. For short, the sigma factor σ_{clip} is defined as a threshold beyond which the value is considered an outlier to be removed from the sample. In this case we used $\sigma_{clip} = 2\sigma_{lqs}$, where σ_{lqs} is the standard deviation of the least-square fit, defined by:

$$\sigma_{lqs} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y(x_i))^2}, \tag{2.2}$$

where y_i is the i^{th} value of the variable to be predicted, x_i is the i^{th} value of the independent variable, $y(x_i)$ is the model predicted value of y_i, and N is the large of the data. After clipping the stars, the linear fit is performed once again in order to obtain the new calibrated photometry m_{cal}, and its error σ_{lqs}.

The final photometric error for individual source was computed using the error propagation:

$$\sigma_{final} = \sqrt{\sigma_{phot}^2 + \sigma_{lqs}^2}. \tag{2.3}$$

The output of the calibration process are individual K_s-band catalogs, in VISTA system, for every available epoch on the tile under analysis. Also the catalogs contain the Modified Julian Day (MJD), taken from the VVV image header.

In the case of the H- and J-band, the procedure, extraction, and calibration in the VISTA system were carried out in a similar process to analyze the d001 and d002 FoV and G 305 region. Later, the VIVA multi-band catalog (Ferreira Lopes et al., 2020) was used for simplicity.

2.1.3 Extracting the VVV time series

All the individual K_s-band catalogs were cross matched (included stilts reference) using a tolerance of 0.4 arc-secs in order to create a multi-epoch master catalog per tile. This master file contains all information of detected sources.

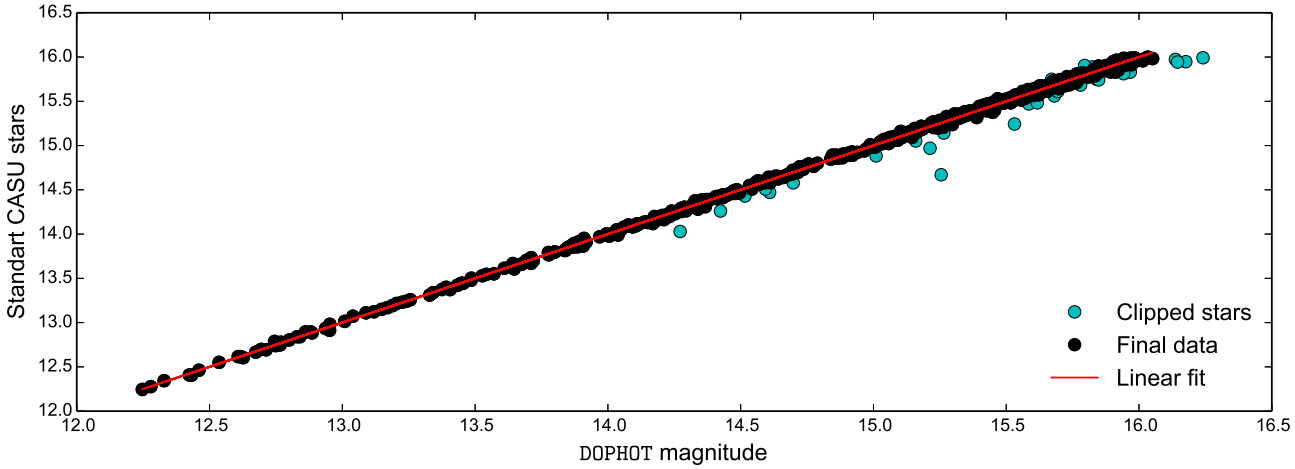


Figure 2.3: Example of the calibration of the DOPHOT photometry with the CASU standard stars, the stars rejected with the sigma-clipping technique are in cyan circles.

Typically this catalog uses several gigabytes of memory and is used to extract the time series using the Universidad de Valparaíso computing cluster.

We filtered our initial sets of time series with some ad-hoc quality and robustness criteria: 1) A minimum of 20 photometric measurements. 2) A total amplitude $\Delta K_s > 0.2$ mag, where $\Delta K_s = \min(m_i) - \max(m_i)$ (more details on section 4.11). 3) An upper limit in flux of $K_s > 11$ mag (to avoid objects that may suffer from saturation in some VVV epochs).

The first restriction represents the minimum number of epochs which allows searching for reliable periods.

The second is motivated by a conservative estimation of photometry errors and transformation to the standard system. The measured uncertainties for the faintest sources were $\sim$ 0.15-0.2 mag for an average K_s at $\sim$16$-$17 mag. This amplitude limit was also intended to decrease the number of sources that were not relevant for a variability study, given the extensive data set obtained from extracting time series from two of the lowest stellar density tiles in the galaxy (tiles d001 and d002).

Figure 2.4 represents the amplitude distribution of the six tiles on the south disc. The black line represents the fraction of the rejected time series with

$\Delta K_s < 0.2$ mag, reaching between 15% and 20% of the total number of sources in some tiles.

Both panels in figure 2.4 represent different methods to display the thickness of the bars in the histograms, the upper panel was basically the initial estimation of the amplitude distribution, and the lower panel displays the identical distributions using the optimized bin-width discussed on section 4.4.1.

These initial filters reduced the analyzed sources numbers obtained from photometry by approximately 30% (for example, from 669825 to 433102 for d001). The total number of extracted sources and for all the considered tiles can be consulted on 3.1.

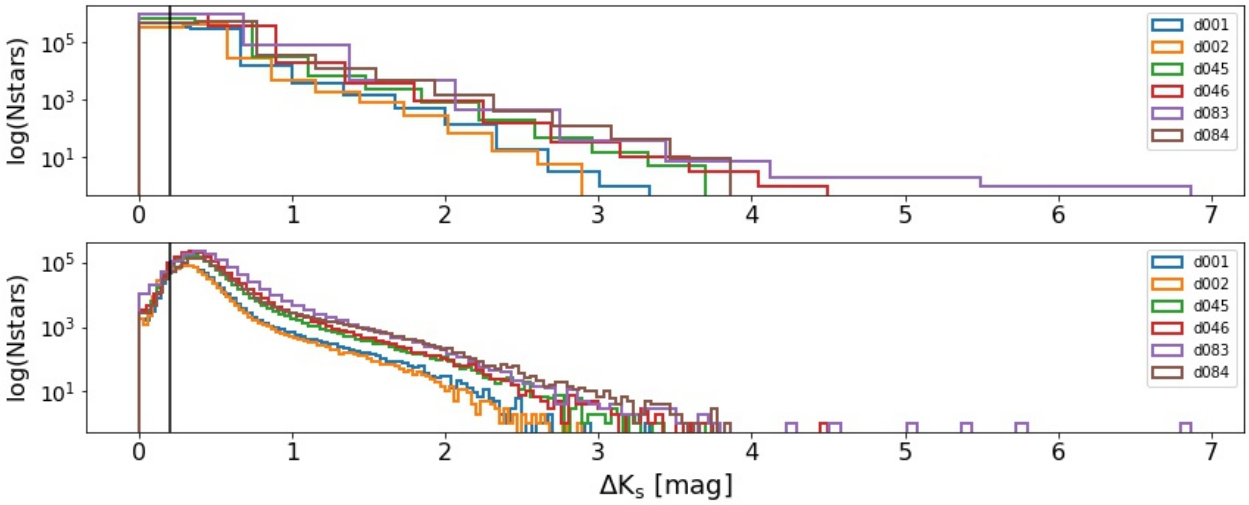


Figure 2.4: The distribution of total amplitude ΔK_s of sources identified inside six tiles that compose the SFR1(tiles d001 and d002) and SFR2 (d045, d046, d083, d084). At the top, the distribution is presented using the default size of bins in the histogram (inside python's matplotlib libraries), where the limit in amplitude $\Delta K_s < 0.2$ mag for the extraction of time series was thought. In the bottom plot, the bin size was estimated using an optimal size considering the method 4.4.1.

Moreover, the photometric measurements are prone to be affected by systematic errors that are hard to clarify and quantify, given atmospheric or instrumental problems. For example Alonso-García et al. (2015) reported a problem related with highly variable PSFs in tile images, due to different geomet-

ric distortions in the combination process of the paw-print images. Thus, to remove outliers in the time series, we implement the modified Thompson τ technique, which is based on the definition shown in Thompson (1985). The modified Thompson τ statistic is defined and discussed on Chapter 4: Methods, in subsection 4.4.2. One of the consequences of this approach is that we will remove poorly sampled transients event from our time series.

A schematic summary of the photometric process explained in this chapter can be seen on the left side of figure 2.5.

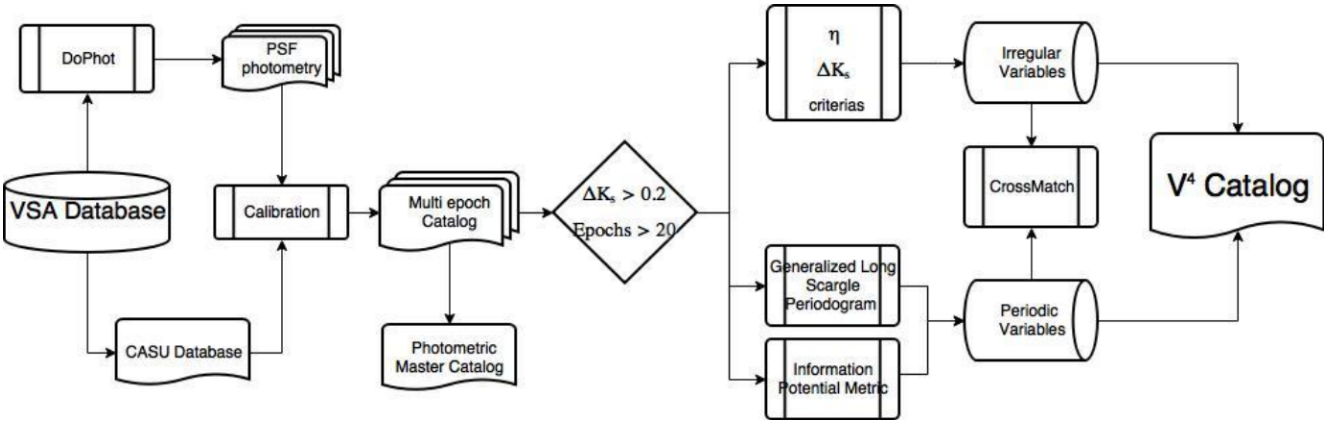


Figure 2.5: Schematic view of the automated process developed and used in this study to categorize the sources in this manuscript.

During our analysis we also used the GAIA data releases DR1, DR2 and DR3 (thought the TOPCAT software (Taylor, 2005)), the IRAC and DSS catalogs. The InfraRred Array Camera (IRAC; (Fazio et al., 2004)) is an infrared camera of the The Spitzer Space Telescope, which operates in four bands: Band 1 (3.6μm), Band 2 (4.5μm), Band 3 (5.8μm) and Band 4 (8μm). The instrument has four detector arrays in the camera, and each has 256 x 256 pixels in size. It is a general-purpose camera that is used by observers for a wide range of astronomical research programs. The data from IRAC were obtained through the ALADIN interactive sky atlas (Bonnarel et al., 2000). The DSS was obtained from Aladin Sky Atlas (Bonnarel et al., 2000). More specifically, the Digital Sky Suvey was used to map and compose the Field of View (FoV) images of the considered SFRs in this Thesis.

CHAPTER 3

Targets

This chapter addresses description and characterization of the selected SFRs chosen to perform a census of variable stars within their vicinity. The focus is to detect and characterize the changing brightness profile of young sources ($<$ 1Mys).

As is commonly observed in spiral galaxies, most of the dust and gas of the Milky Way lies within a thin disk of a few hundreds of parsecs (approx 500 parsecs, reference). This is the place where interstellar gas is located. Is composed primarily of Hydrogen and Helium which comes from the Big bang.

To perform a census of YSOs it is necessary to explore the densest regions within the ISM where the main initial condition for star formation is fulfilled: High concentrations of hydrogen. In general, the YSOs are expected to be correlated with the overdensities of molecular hydrogen H_2, commonly called molecular clouds, composed by atomic hydrogen (HI) surrounded by ionised hydrogen (HII).

The ISM have progressive enrichment with Helium and heavier elements. This enrichment give different scenarios inside the stellar evolution. For example, the abundance of heaviest elements inside the ISM depends on the distance of the center of the galaxy, wherein the very center the abundances are

larger than the edge of the thin disk can be measured using the metallicities of association of stars and clusters. We are interested in exploring the diverse SFR along with the bulge and the south disk of the galaxy in order to explore environment dependence (if any). The targets and basic information of the selected regions can be seen on table 3.1.

3.1 Cadence of the VVV observations

The VVV observation strategy was focused on covering a large area on the sky, using the VISTA tile field of view is 1.501 deg2. Considering the complete area to observe, 196 independent tile region where needed to map the bulge area and 152 tiles for the south disc. Figure 2.1 provides a schematic representation of the executed scheme for the Galactic center region and the south disk region. The full survey required a total of 192 nights of observations over 5 years (Minniti et al., 2010).

Such strategy leaded to different cadence and number of observations of every individual tile between 2010-2015. The disk fields generally have up to 60 epochs, meanwhile tiles projected to the bulge have between 75-90 epochs. The only exception of the region of Baades' Window, where given the low extinction, there are 8 tiles with more than 300 epochs.

Nevertheless, this cadence of the observations is able to detect variability related to timescale accretion variations, star spots, episodic accretion events, rotational modulation and variable extinction in the YSOs (Contreras Peña et al., 2017b; Rebull et al., 2014a).

For example, the tiles d001 and d002 accumulate up to 51 and 48 epochs observed respectively. Left side of Figure 3.1 shows the gaps, the baseline, and the maximum size of the time-step between epochs. The right side of Figure 3.1 shows the distribution of the difference of consecutive observations ΔMJD, zoomed up to $\Delta MJD > 12$ days. The minimum time interval between the observation is $\sim$0.35 days, with distribution maximum between 0.35 and 2 days.

On the other hand, VVV produces unevenly spaced time series, which in-

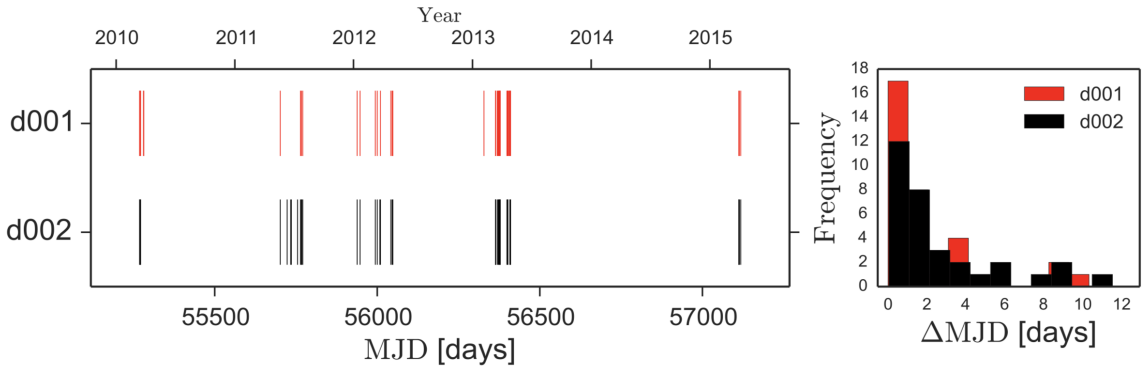


Figure 3.1: Overview of complete observation set used to analyze VVV tiles d001 and d002 taken from Medina et al. (2018). Left side: The log of observations of VVV tiles d001 and d002 between 2010 and 2015, based on the log representation of Rebull et al. (2014a). Each photometric measurement is marked by a '|' symbol. The bars are thicker in places with high cadence. Right side: Histogram of differences between consecutive observations ΔMJD. The typical time interval between observations is between 0.3 and 2 days.

duces uncertainty and a challenging period estimation. Still, we are able to identify different types of periodic variability in a wide range of time scales proportional to the total window of observation (see for example Elorrieta et al. (2016); Gran et al. (2016); Minniti et al. (2017)).

The figure 3.2 shows all analyzed tiles in this PhD thesis, as well as the photometric log summary given in Table 3.1.

For simplicity, the different fields considered will be referred to simply as SFR1 to SFR4, in the proximity to the Galactic center. The position and designation for each region are displayed in the figure 3.3.

3.2 Tiles d001 and d002: IC 2944/2948

The tiles named d001 and d002 were used as a starting fields in order to develop and test the automated process described in Medina et al. (2018) (see also the scheme 2.5 for more details). With more than a million point sources and over 900 thousand time series we also test the quality of the photometry,

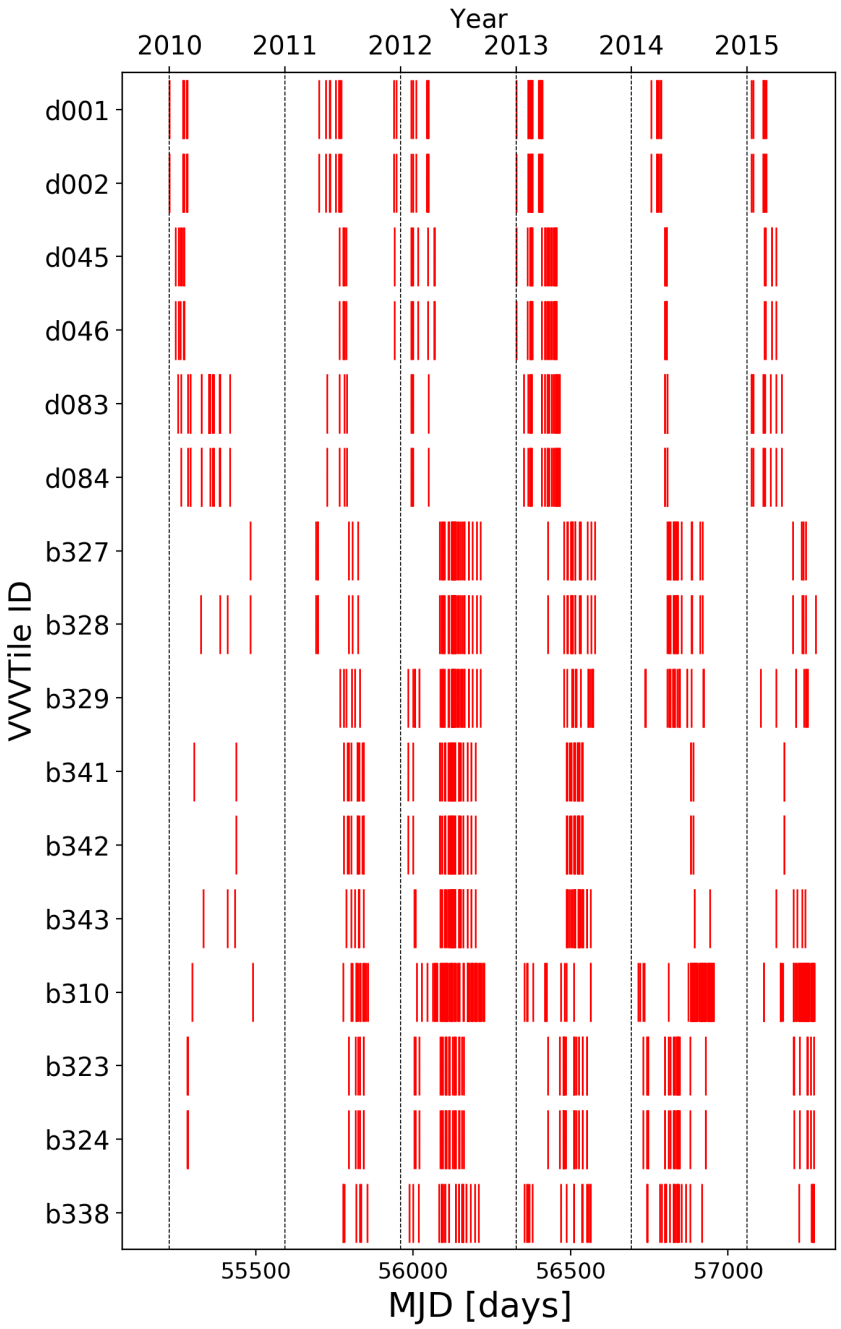


Figure 3.2: The log of observations of VVV tiles between 2010 and 2015, based on the log representation of (Rebull et al., 2014a). Each photometric measurement is marked by a '|' symbol. The bars are thicker in places with high cadence. it is possible to see the different seasons of observation between the disk (summertime at south hemisphere) and the bulge (winter nights at south hemisphere). Also is been included one high-cadence field inside the Baade's window: tile b310, with more than 300 epochs included in their dataset.

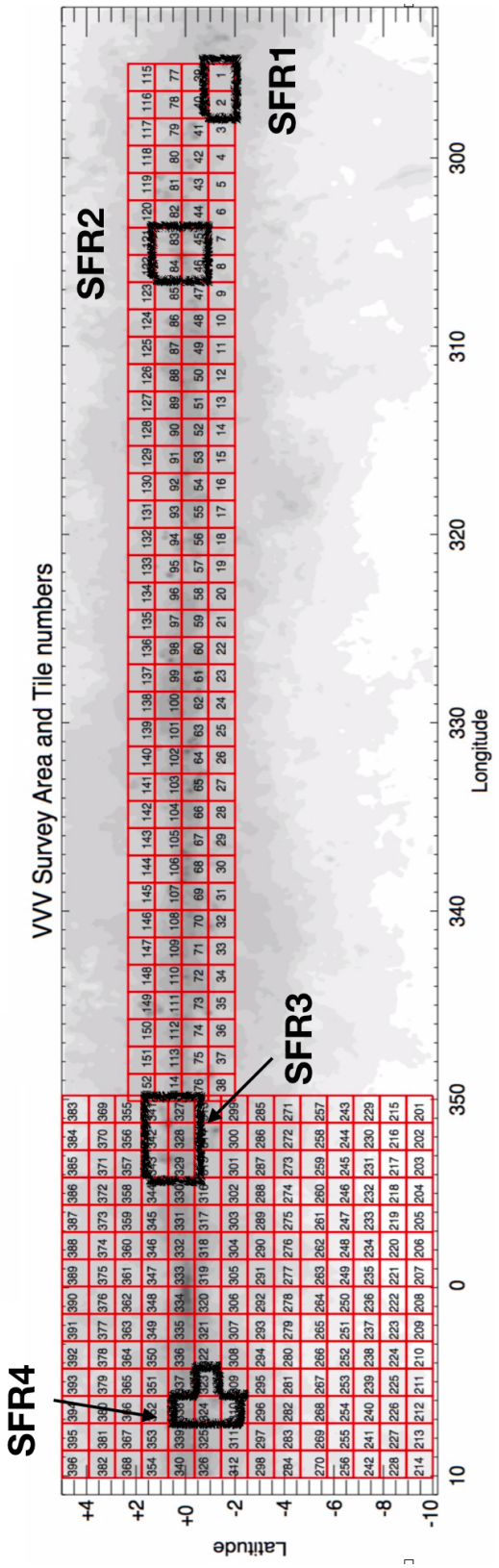


Figure 3.3: The VVV survey FoV with the different Star-forming regions considered inside this study are indicated. The number of epochs inside the SFRs depends on the Galactic longitude.

VVV tile ID	Associated to	Coordinates (ℓ [°], b [°])	Epochs	Sources with time-series	Irregular sources
d001	IC 2944/2948	(295.4,-1.6)	51	433,102	45
d002		(296.8,-1.6)	48	465,634	26
d045	G305	(304.2,-0.5)	55	950,085	57
d046	G305	(305.7,-0.5)	54	1,068,806	66
d083	G305	(304.2,0.5)	74	754,122	32
d084	G305	(305.7,0.5)	59	797,051	41
b327	NGC 6334	(350.7,+0.1)	79	1,956,310	83
b328	NGC 6334-6357	(352.2,+0.1)	83	1,843,695	94
b329	NGC 6357	(353.6,+0.1)	85	2,348,086	57
b341	NGC 6334	(350.7,+1.2)	88	1,309,496	156
b342	NGC 6334-6357	(352.2,+1.2)	87	1,465,298	60
b343	NGC 6357	(353.6,+1.2)	85	1,809,178	106
b310	Simeis 188	(6.7,-2.1)	307	2,332,512	57
b323	M8	(5,3,-0.9)	79	2,389,609	89
b324	M8	(6.7,-0.9)	79	2,233,543	71
b338	Trifid Nebula	(6.7,+0.1)	78	2,452,234	56

Table 3.1: Summary of the VVV observations, sources with performed PSF photometry and selected irregular variables in each tile. The total number of reported Irregular Variable Sources are 1096 objects.

the calibration and the cross match processes.

These tiles are selected because of their low stellar density (in comparison with other areas of the disk, see table 3.1) and the clearly outlined nebulosity distribution in the FoV (see figure 3.4). In the FoV of tile d001 (central coordinates) the open cluster IC 2944 is projected. This is a group of massive O and B stars that excite their associated reflection nebula, projected at the inner edge of the Carina spiral arm.

The nebulosity that surrounds IC 2944 is know as IC 2948 or λ Centauri Nebula (given their projected proximity to Centauri constellation, but is not related). Also, the nebula is know as ”Running chicken nebula”, an strange

nickname from the bird-like shape of its brightest region[1].

Several H II regions are associated to the open cluster IC 2944. For several decades, objects IC 2944 and IC 2948 have been considered as a single very extensive H II region (Thackeray & Wesselink, 1965). The published distance to the IC 2944 H II region varies from 1.8 kpc (McSwain & Gies, 2005) to 2 kpc (Sana et al., 2011). Some part of the nebula IC 2948 falls outside the d001 FoV, particularly the regions with galactic longitude $\ell < 293^{\circ}$ [deg], which is the case of the associated object IC 2872.

Other indicator of the recent stellar formation in the vicinity of IC 2944 is the presence of Bok globules. These relatively isolated and small extinction patches are only visible for the lacking of background luminosity in visual wavelengths. In this particular case, the Bok globules are known as "Thackeray globules" Reipurth et al. (1997) and are connected with low mass star formation.

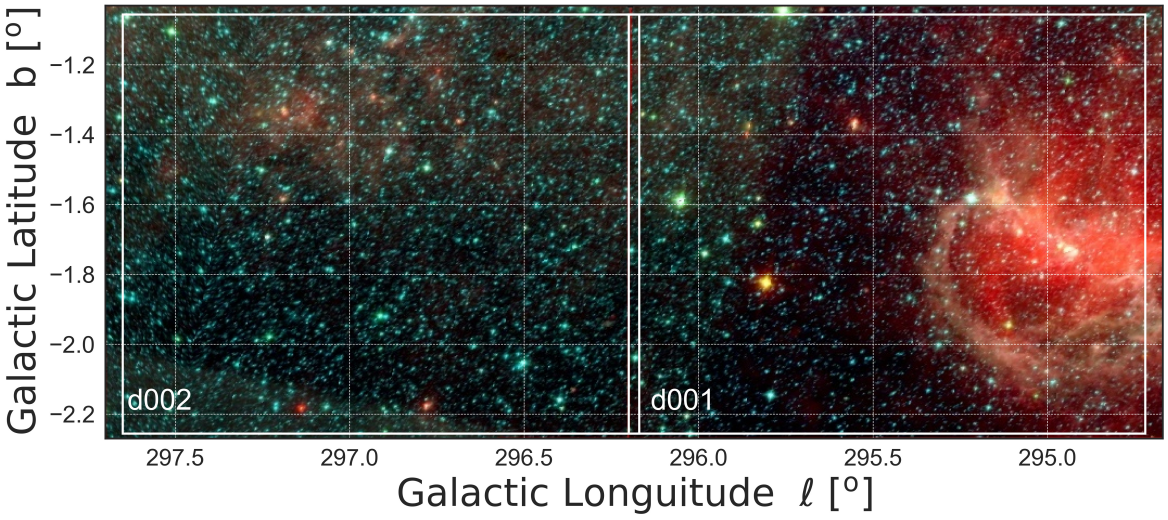


Figure 3.4: Covered region by d001 and d002 tiles from VVV survey, showing the spatial distribution of identified sources. In the background, the ALLWISE RGB false-color image using the W4 in red ($\lambda_{eff} = 22\mu m$), W2 in green ($\lambda_{eff} = 4.6\mu m$), and W1 in blue ($\lambda_{eff} = 3.4\mu m$) is shown to illustrate the cold gas/dust distribution in the FoV. Galactic north is up; galactic east is to the left.

[1] https://www.eso.org/public/images/eso1135a/

3.3 The G305 Star-forming region

G305 is one of the most massive star-forming regions in our Galaxy, located within the Scutum-Crux arm (in galactic coordinates: $l = 305.506°$, $b = 0.085°$). Its integrated radio luminosity can be directly comparable to the most luminous giant H II regions of the Galaxy (Conti & Crowther, 2004). G305 is also a valuable laboratory for studying sequential star formation. According to Clark & Porter (2004), there are at least two different generations of young stars within the G305 radio complex.

The first generation is associated with the massive clusters Danks 1 and 2, which contain many massive OB and Wolf-Rayet stars. The well known Wolf-Rayet star WR48a is projected nearby. The presence of a dusty WCL star implies a minimum age of $\sim$3 Myr and the lack of the red supergiants in the region suggests a maximum age of $\sim$4-5 Myr. The second generation consists of embedded IR sources, maser emission and ultracompact H II regions (UCHIIs) (Hindson et al., 2012). A significant portion of the radio and submilimetric emission comes from the regions which coincide with the periphery of the mid-IR nebula (Clark & Porter, 2004).

In the central cavity are located two young and massive star clusters: Danks 1 and Danks 2. The molecular cloud from these clusters has dissipated, but molecular clouds still exist around the cavity sculpted by those young clusters.

The lack of this molecular material around implies that star-formation has ceased in Danks 1 and 2, but these nearby regions may be sites of ongoing star-formation. Several H II regions and YSOs are found in the dense molecular material surrounding the two clusters. This H II regions are understood to be ignite by massive stars with strong stellar winds. Hence the star formations is driven by the presence of massive stars, where the strong stellar winds and radiation from such a population can provide the enough energy to trigger stellar formation.

One of the panoramic Infrared views of G305 star-forming complex is shown in figure 3.5. The false-color view provide by the Spitzer Space Telescope mid-infared InfraRed Array Camera (IRAC) Fazio et al. (2004) at wavelengths of 3.6 to 8 μ m is in background. Visible light images reveal no trace

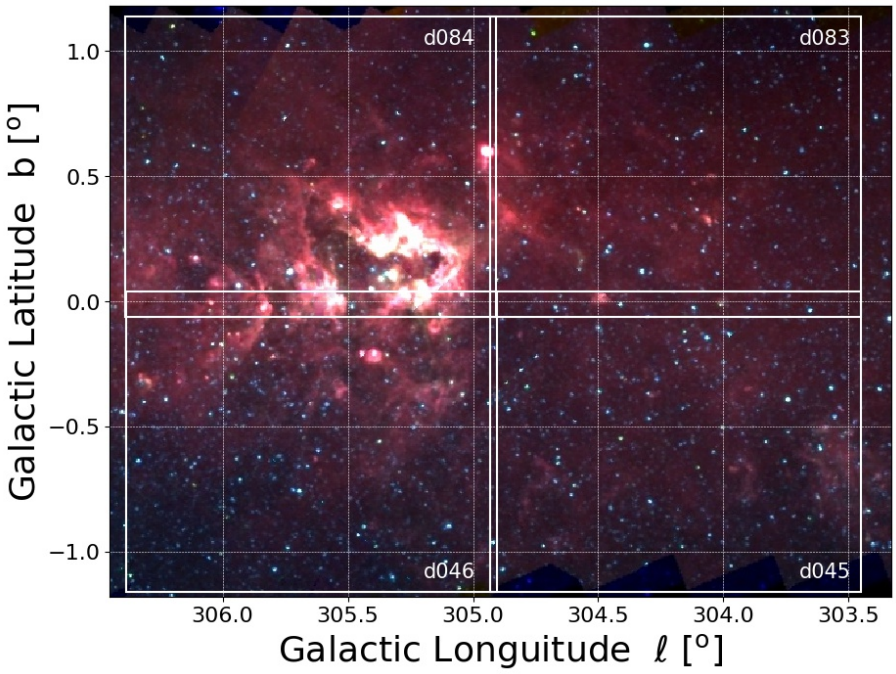


Figure 3.5: The G 305 star-forming complex region. The four observed tiles, named d045, d046, d083, and d084 from the VVV survey are labeled. In the background, the false three-color GLIMPSE Infrared Array Camera (IRAC) (Fazio et al., 2004) image of the complete field is shown, using 8.0μm (red), 4.5μm (green), and 3.6μm (blue). Galactic north is up, galactic east is to the left.

of what is happening in this region, the heavy dust obscuration hide all the radiation that came from the new generations of stars.

The complex pattern of filaments that shape the interstellar medium is caused by stellar winds of massive stars. Also the presence of polycyclic aromatic hydrocarbons (PAH) (luminescent at wavelengths near 8.0 microns) trace the cold molecular gas and the outflows from the new generations of sources within the natal cloud.

Great part of molecular content revealed by Spitzer telescope inside this SFR is covered by VVV tiles d046 and d084. Surrounding areas of G305 SFR were also considered and covered by VVV tiles d045 and d083. There are many signs of active star formation, invisible for IRAC cameras.

3.4 NGC 6357 and NGC 6334

The active high-mass star-forming complexes NGC 6334 and NGC 6357 are two extensive giant molecular clouds located in the Carina–Sagittarius Arm of the Milky Way, less than 10 degrees from the very center of the galaxy (see, e.g., Chibueze et al. (2014)). Both regions are connected by a filamentary structure $\sim$2 degree away to the north of the galactic plane, so it was proposed that both areas may belong to a single and extended complex (Russeil et al., 2010, 2017). More still, the complexity and the presence of a large number of O stars, the gas/dust condensations, and the mixture of high and low-mass star formation of this region are also explained, proposing a "Galactic mini-starbursts" Fukui et al. (2018). Several observations in infrared domains had been executed in the vicinities of these massive star-forming complexes for a better understanding and characterization of the environments.

The visual morphologies of both molecular clouds (observed in figure 3.6) are dominated by almost spherical H II cavities shaped by the radiation of close massive young stars. In the case of NGC 6357, the open cluster Pismis 24 is located at the center of the molecular cloud, in a similar way to Danks 1 and 2 inside G305 SFR complex (section 3.3). On the other hand, NGC 6334 displays groups of H II regions scattered around the central filament of molec-

ular hydrogen and dust. Features of heavy extinction are also present in this Infrared Dark Clouds (IRDC), which is also known as the "star-forming axis" (Reipurth, 2008), so the probability of the presence of embedded, undiscovered massive (and low mass) stars is largely due the dramatic interstellar extinction inside the FoV.

Different methods for distance estimation have been used regarding the distances of NGC 6334 and NGC 6357. Inside the review (Reipurth, 2008), NGC 6334 is located at a distance of 1.61 $\pm$ 0.08 kpc, using the average value for the eleven OB stars that ionized the optical H II nebulae. Meanwhile, in Lima et al. (2014), using the VVV survey, identified five embedded clusters within NGC 6357 and derived a distance of 1.78 $\pm$ 0.1 kpc using the color-magnitude diagrams.

The most considerable discrepancies are published in Chibueze et al. (2014), where the parallax of H2O maser sources inside a dense molecular clump associated with NGC 6334 was used for estimate distances of 1.35$\pm$0.15 and 1.25$\pm$0.14 kpc, respectively. It seems now widely accepted that NGC 6334 and NGC 6357 are both located at a distance of $\sim$1.7 kpc (Fukui et al., 2018).

Six VVV tiles cover the entire field of both NGC 6334 and NGC 6357 and their intermediate environments: b341, b342, b343, b329, b328, and b327. Given these tiles are projected in the bulge region, individual tiles have at least 79 epochs each, as can be seen in table 3.1.

3.4.1 NGC 6334

NGC 6334 ($l = 351.2°, b = +0.7°$) is popularly known as "Cat's Paw Nebula" given the effect of the presence of OBs stars in the optical emission nebula. Several OB stars lie around the high-mass filamentary cloud structure, with an estimated spanning of 12 pc. Given the dimensions of this star-forming complex, several embedded star-forming regions have been identified from infrared and radio emissions.

The first five main Far-infrared dense sources received the names NGC 6334-I to NGC 6334-V. Later, other Far IR sources named NGC 6334-I(N) were identified, associated with HII or UCHIIs regions. Optically, NGC 6334 is

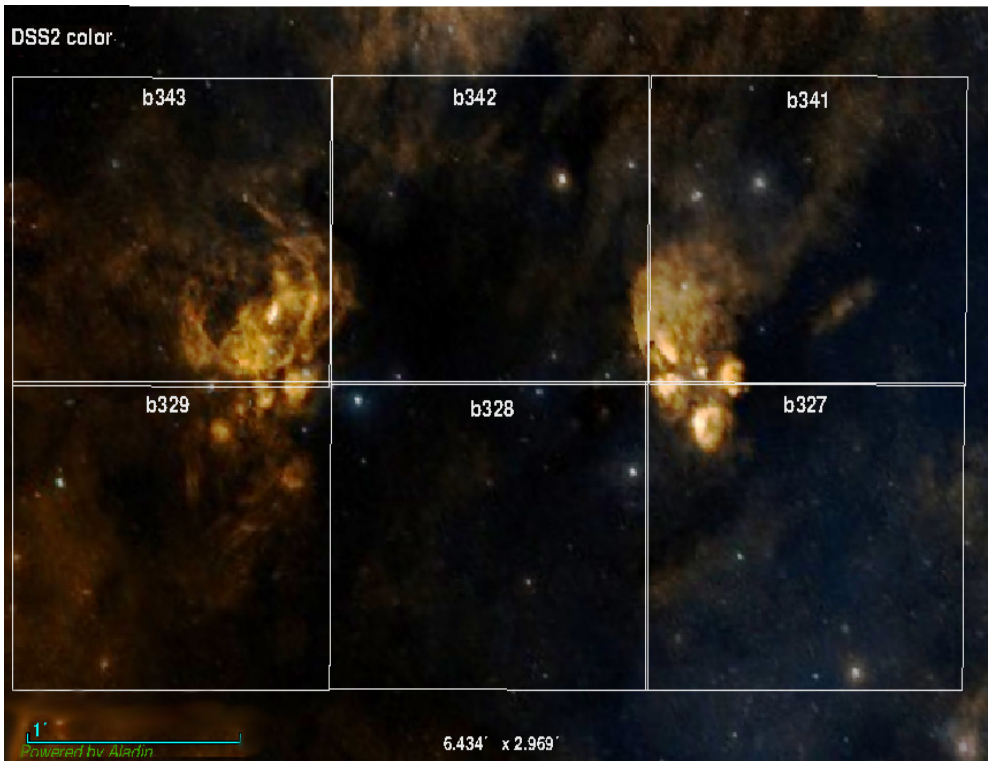


Figure 3.6: A panoramic image of the largest FoV considered inside this manuscript. The visual nebulosities of the Cat's Paw Nebula (NGC 6334) in on the right, and the Lobster Nebula (NGC 6357) is on the left. In the background, the DSS RGB image of the considered FoV is displayed. In white, the footprints of VVV tiles b341, b342, b343, b329, b328, and b327. Galactic north is up; galactic east is to the left.

a grouping of the well-known optical HII regions, called GUM61, GUM62, GUM63, GUM64, HII351.2+0.5 and GM1-24, and presenting more compact H II regions than NGC 6357 Russeil et al. (2016). Some of these shell structures are considered as examples of the so-called Strömgren spheres, as the case of GUM61.

In the past few years, this SFR has been selected for the study of their monolithic filaments properties and individual star-forming regions at large scales using radio continuum observations. Also, there are four main clusters along the so-called star-forming axis: NGC 6334-I, NGC 6334-I(N), NGC 6334-I(NW), and NGC 6334-E; this last one was discovered using the continuum dust emission from dense stellar cores with ALMA (Sadaghiani et al., 2020). Among these star clusters, NGC 6334-I(N) is the most populated cluster, with 80 cores.

VVV tiles b341 and b327 cover almost wholly the "Cat's Paw" emission nebula and a significant part of their star-forming complex. Tiles b342 and b328, on the other hand, will cover the region that surrounds NGC 6443 and the area that presents the connection with NGC 6357.

3.4.2 NGC 6357

NGC 6357 ($l = 353.16°, b = +00.89°$) is also know as "The Lobster Nebula" is two deegrees closer to the very galactic center than NGC 6334 at the same galactic latitude l. The morphology of NGC 6357 is also dominated by the presence of spherical H II regions produced by massive stars. The presence of a large central cavity is observed in visual wavelength and several shell-structured shapes around the central cavity, where the massive open cluster Pismis 24 is located, which contains a few of the most massive stars known. In particular, Pismis 24-1 (or HD 319718) is a multiple O star system inside the cluster AH03J1525-34.4 (Fukui et al., 2018). YSOs population has been observed concentrated in the surroundings of Pismis 24.

The brightest HII region is called G353.2+0.9, and it traces the sharp boundary of Pismis 24. Some small, optical HII regions can be observed in the surroundings, named H II 353.42+0.45, 353.08+0.28, 353.09+0.63, and 353.24+0.60.

Infrared surveys detected several sources with NIR excess projected into the nebulosities. Radio-telescopes also display the presence of compact sources projected into the nebulosities.

VVV tiles b343 and b329 cover the ”Lobster” emission nebula.

3.5 The Lagoon nebula complex

The Lagoon Nebula complex is a massive star-forming which gathers several different and independent massive HII regions, emission nebulae, low-extinction regions, and several types of astronomical objects inside their ISM, as supernovae remanents as W 28 (e.g., Green (2014)) and planetary nebulae (e.g., PN M 1-41, Velghe (1957)).

This complex region is located on the Sagittarius-Carina Arm, lying a few degrees away from the line of sight of the Galactic center.

As is described inside the review (Tothill et al., 2008), this complex has been observed and cataloged since very early in history, particularly from the 17th century, therefore exist several nomenclatures for identifying stellar objects. In this manuscript, the ”Lagoon nebula complex” name is used to refer to the group of optic nebulas called M8, M20, and Simeis 188 and some degrees around their optical nebulas.

Figure 3.7 visualizes FoV covered around the Lagoon Nebula complex. The footprint of selected VVV tiles that cover the region is displayed in white. The central HII region and most extensive nebula inside the complex: M8 (005.9575 -01.1667), with an estimated radius of about ten pc. Using M8 as a reference, the Lagoon Nebula complex also includes the optically visible M20 (007.0859 -00.2873, a.k.a. Trifid Nebula), located in the galactic northwest. The optical-invisible HII region W28 A2 (a.k.a. G 5.89-0.39) is situated to the north; finally, the complex of nebulae to the east of M8, known as Simeis 188, is located one degree to the west.

Tile b338 contains The Trifid Nebula and its surroundings. Tiles b323 and b324 cover M8 and their surroundings as An essential part of the ultra-compact HII region W28 A2 SFR, which falls inside the FoV of tile b324, despite that the

central area of the W28 A2 is not within their FoV. Finally, tile b310 covers the several reflection nebulae projected in front of M8, called Simeis 188, a collection of spherical HII regions, bright emission nebulae, and IRDCs.

3.5.1 The Lagoon nebula

M8 ($l = 005.9575°$, $b = -1.1667°$) is one of the largest and bright HII regions known in the galaxy. M8 is composed of four main structures: The young cluster NGC 6530 ($l = 6.082°$, $b = -1.331°$), the M8 E star-forming region ($l = 6.05°$, $b = -01.449°$), The Hourglass Nebula ($l = 5.9736°$, $b = -1.175°$) and the HII region NGC 6523/33 which is the nebula itself.

Several features of stellar formation surround the visual nebulosities of NGC 6523/33 as open clusters, numerous bright rims, bok globules, cold molecular pillars (a.k.a. elephant trunks), and IRDCs. (Lada et al., 1976) identified that the star Herschel 36 ($l = 5.970°$, $b = -1.173°$) is the primary source of ionization in the region. Also, the sources 9 Sgr (6.008 -01.204, an O4V(f) star) and HD 165052 ($l = 6.121°$, $b = -1.481°$) contributes to the ionized gas inside the ISM (Tothill et al., 2008).

In the front of the interstellar material of NGC 6523/33, close to the nebula's center, lies the massive open cluster NGC 6530, which has been determined to be extremely young. With a large population of pre-main-sequence sources and several O and B stars, the average estimated age is below one Myr Prisinzano et al. (2019), and a distance of approximately 1,32 Kpc Damiani et al. (2019).

The Hourglass Nebula and the M8-E H II regions are located on the galactic west and east of the main nebulosity NGC 6523/33, respectively. Both are massive star-forming regions powered by several O-type stars. Several different populations of young stellar populations have been detected using IR and X-rays wavelengths.

The distribution of the very young classical T Tauri stars with reddening excess correlated to the position of the O-type stars can display evidence of triggered star formation by the O-star ionization fronts (Prisinzano et al., 2019); and hence, at least a second generation of stars.

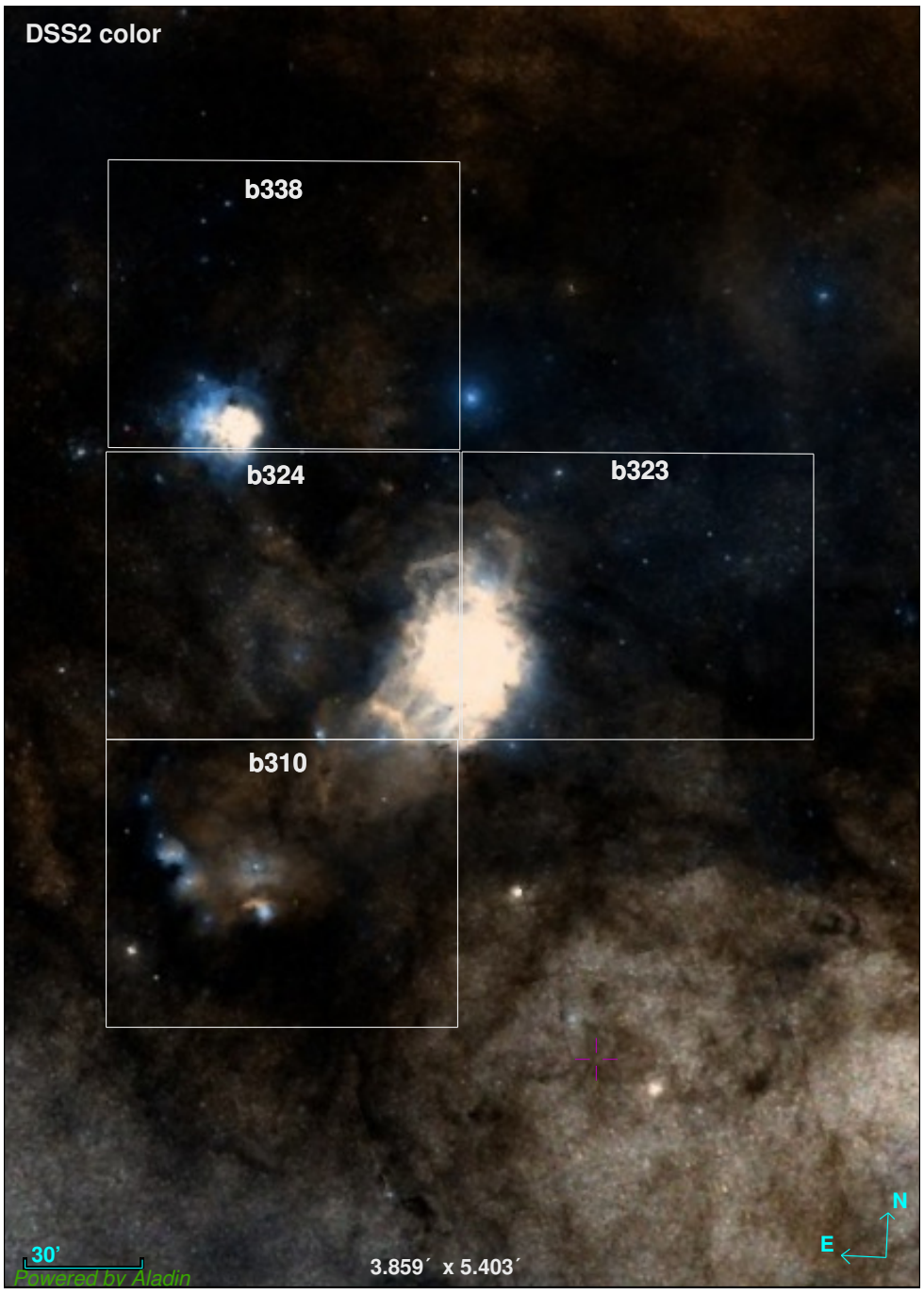


Figure 3.7: DSS optical RGB image of the considered field that contains The Lagoon Nebula Complex. The Trifid Nebula (M20) is the northern nebulosity and falls into a small region of the tile b338. The Lagoon Nebula (M8) is covered by tiles b324 and b323. The Simeis 188 region is projected over the high-cadence VVV tile b310. The footprint of VVV tiles are marked in white. Galactic north is up; galactic east is to the left.

3.5.2 The Trifid Nebula and their surroundings

The Trifid Nebula (M20) is a young H II region with multiple sites of active star formation. With a minor diameter of only 3.5 pc, the numerous presence of O star, and the subsequent young stellar generations, including the current protostellar population forming now, in and around the nebula (Rho et al., 2008).

t comprises two visual nebulas: in the north lies the reflection nebula surrounding the A7 supergiant HD 164514, and the emission nebula entirely powered by a young massive star O7 named HD 164492. Many cores exhibit signs of dynamical interaction with the ambient medium. Therefore M20 is believed to be formed from the dynamic impact of HD 164492 in a turbulent environment.

A large sample of YSOs has been detected using color-color diagrams, X-ray emission, and CO emission maps. The star HD 164492 is surrounded by a cluster of approximately 3100 young stars; among them, five stars have been early identified within 20´´ of the central O star. Kohoutek et al. (1999) confirmed that the brightest components, HD 164492B to HD 164492B-E, were early-type stars forming a small stellar grouping. Also, Four massive (25–90 solar masses) protostellar cores (TC1 to TC4) were discovered in the Trifid with millimeter observations (Cernicharo et al., 1998). Compact H II regions such as G8.1+0.2 and G7.5+0.1, and young stars undergoing collapse and violent mass ejection present more evidence of massive star-formation in the vicinities.

It was thought to lie between a range of 1.6-2.7 Kpcs away; nevertheless, parallaxes from Gaia EDR3 have been determined to be around 1.25 Kpcs, placing the region in the Sagittarius arm of the Milky Way (Kalari, 2021). The Trifid Nebula falls in a small region of the VVV tile b338, so an extensive area to the galactic north where several IRDCs are distributed and several signatures of active star formations are located between the dust filaments. Finally, the Trifid Nebula is famous for being the cover of the album ”Islands” by the band King Crimson.

3.5.3 Simeis 188

The region known as Simeis 188 (Herbig 1957) is a collection of visual emission nebulae: (IC 1274 and IC 1275), reflection nebulas (IC 4684), IRDC (B 91, B302, and B303), and the SFR known as NGC 6559, also known as a bright-rimmed cloud 89 (BRC89, Ogura et al. (2002)). These structures seem to be connected by a dense filament of dust detected in visual and IR wavelengths. Different O-type stars ignite these emission nebulae. Simeis 188 lies close to the low-extinction region known as the "Baade's window" (Baade, 1946) and is of great scientific interest due to its properties. VVV tile b310 covers the elements that compose Simeis 188 and is the FoV with the most significant number of observations between 2010 and 2015, rising up to 300 epochs.

Chapter 4

Analysis methods

This chapter will focus on describing the mathematical methods, the data transformations, and the algorithms that are implemented to analyze the uneven-sampled VVV time series.

There are different approaches to directly extract information from time series, such as methods to transform, summarize, reduce and describe data. In general, astronomy has always been an excellent source of information for developing new technique analyses and technologies. Still, the growing interest in the structure and properties of time series is not only in astronomy. Several fields in science use similar tools for this purpose and share their results with the community.

We will concentrate on the time-domain data and will describe and construct a set of statistics that quantify some valuable aspects from the time series generated by the measured stellar magnitude over time. The analysis of the so-called "color-color diagrams" (CCD) will also be discussed and included to obtain relevant conclusions about the nature of the variable source.

4.1 Variability indicators in time series

With more than 25 million detected stars, every time series $\{x_i\}$ contains between 20 and 300 observations depending on the galactic longitude $\ell[^\circ]$. A single file for every $\{x_i\}$ was created with three data columns: the MJD time t, the K_s-band magnitude m, and the photometric uncertainty σ. This set of data can be represented as follows:

$$\{x_i\} = \{\mathrm{t}_i, m_i, \sigma_i\}, \tag{4.1}$$

where i is the total number of epochs inside the time series. The current analysis is based on the dataset $\{x_i\}$ where, which, as discussed in section 1.4.1 describes a time series. Nevertheless, the use of time series in astronomy is not something new; there were papers decades ago where commonly, the sequential measurements of a particular star were referred to as a series of photometric observations as m_i.

Under this notation, $m_i = m(t_i)$. It is important and interesting to see implicit time dependence in this definition. Also, somehow, it indicates the reductive nature of time series analysis, where more available information is requested to define a specific statistic. For example, the classic definition of the arithmetic mean value of the mean magnitude K_s follows:

$$\overline{m}(\{x_i\}) = \frac{1}{N}\left(\sum_{i=1}^{N} m_i\right). \tag{4.2}$$

As can be seen, the input of the statistic "mean" is the time series $\{x_i\}$, but not all of the elements inside the data set are used. Only the photometric magnitudes m_i were useful to obtain $\overline{m}$. Not all the time series components will necessarily be used to define a new measurement or statistic. Even so, exists statistics that use all the components inside $\{x_i\}$ and even define more quantities to complement and put together within a mathematical expression designed to obtain some specific information from one particular star in the galaxy.

The mean defined in 4.2 itself quantifies the central value of a set of magnitudes and is a vital and basic piece of information for any stellar source. Nevertheless, in terms of variability detection, the arithmetic mean is pointless. A better option to quantify the variability of a parameter could be the standard deviation, which is a measure of dispersion or the characterization of the scatter of a set of values:

$$\sigma_{ts}(\{x_i\}) = \sqrt{\frac{1}{(N-1)}\sum_{i}^{N}(m_i - \overline{m})^2}, \tag{4.3}$$

The standard deviation quantifies how far the data points lie from the mean value.

In astronomy, it is frequent to use different sets of statistics (Sokolovsky et al., 2017), commonly called "variability indices" or simply "features", to quantify changes in luminosity with time. Depending on the definition of these indices, a population of variables that have similar behavior can be identified, and one can try to separate stochastic variability from "further organized" flux variations. Classic examples of these indexes are the Welch-Stetson I_{ws} (Welch & Stetson, 1993); Stetson J_{Stet} and K_{Stet} indexes (Stetson, 1996). These quantities have been used to identify sources that exhibit large photometric variations over time. Defined in this manner, sources with larger J_{stet} values are the most probable variable sources. Different authors in the literature define particular limit values of J_{stet} for this task (such as $J_{stet} \geq 0.55$ (Carpenter et al., 2001), $J_{stet} \geq 0.9$ (Rebull et al., 2014a)). In the literature, it is possible to find many more different features tailored to identify different types of variable sources. Thus, deciding which index is useful to detect a specific type of variability depends not only on the definition of the index itself but also on the properties of the available data.

4.2 ¨Feature¨ selection

The next section will describe the set of features that we applied and used to quantify the flux variation in different variable star sources in selected SFRs en-

vironments across the galaxy. The idea is to describe the overwhelmed quantity of information in a set of data that represents every time series $\{x_i\}$ using an ad-hoc set of statistics or features chosen among all the possible definitions in the literature. The features were selected considering the characteristic of the VVV time series: a single band time series have taken on unevenly sampled observations with a different observational log for each considerate tile (see a visual summary of the observation for every tile in 3.2).

Following the discussion about the features in the literature, the tendency is to apply robust variability indices to avoid the possible scattering that the unevenly sampled observations can introduce. Also, it is common to use a broad selection of variability indices (and mathematical definitions) to cover different scenarios and procedures that can produce brightness changes.

Considering the information from literature and the properties of the VVV time series, a set of seven features has been selected to characterize the behavior of the stellar sources over time. They are summarized in Table 4.1.

Index	Reference
Stetson J	Stetson (1996)
Stetson K	Stetson (1996)
η index	von Neumann (1941a)
Total amplitude ΔK_s	Contreras Peña et al. (2017b)
σ/μ ratio	Shin et al. (2009)
Classical χ^2 test	Rebull et al. (2014a)
Period P	Lomb (1976); Huijse et al. (2011)

Table 4.1: Set of Features or variability indices used in this analysis and the references where the features are defined.

4.2.1 Stetson J_{Stet} and K_{Stet} indexes

The *Welch-Stetson index* I_{ws} was proposed by Welch & Stetson (1993) during the Cepheid variable stars study observed by the Hubble Space Telescope (HST). It was designed to combine N "almost simultaneous" independent observations

for multi-band photometric measurements of the same source, thus connecting the changes in the color with the intrinsic variability.

The fact that more than one photometric time-series have to be considered needs the definition of normalized magnitude residuals for a given band-pass and scaled by the standard error $\delta m_i = \left(\frac{m_i - \overline{m}}{\sigma_i}\right)$. Thus, both (or more) time-series can be directly combine in one statistic. Furthermore, Peter Stetson extended the definition for the normalized magnitude residuals adding the bias factor, namely the relative error for the i-th term δ_i is defined as:

$$\delta_i = \sqrt{\frac{N}{N-1}} \left(\frac{m_i - \overline{m}_w}{\sigma_i}\right), \tag{4.4}$$

where σ_i are the uncertainty of the individual measurements m_i, and $\overline{m}_w$ represents the weighted average magnitude, considering the weights as the inverse of the photometric uncertainties $w_i = (1/\sigma_i^2)$:

$$\overline{m}_w(\{x_i\}) = \frac{1}{N} \frac{\sum_i^N w_i m_i}{\sum_i^N w_i}. \tag{4.5}$$

Taking advantage of the definition of δ_i, Stetson proposes to relate two or more observations in several filters taken in short periods. Therefore he proposed to create an expression that divided the observation time interval into several convenient binned slots k. If two (or more) time series $\{x_i\}$ and $\{x_j\}$ were obtained from the same source inside the same temporal bin k. This statistic is so-called the "*pair* P_k".

One of the most significant disadvantages of the I_{ws} index is the assumption to have simultaneous observation pairs in more than one or several photometric filters. Therefore, three years later, (Stetson, 1996) updated the statistic and included the option to evaluate single observations from a single time-series in the pair P_k definition, which is the following:

$$P_k = \begin{cases} (\delta_i \delta_j)_k, \text{ for paired k observations, with } i \neq j \\ \delta_k^2 - 1, \text{ for single observations} \end{cases} \tag{4.6}$$

This last addition to the pair P_k represents the unbiased measurement of δ_i, due

the expectation value $\langle \delta_i^2 \rangle = 1$. After bringing all this information together, it is possible to define "a more robust version" of the index I_{was}: the *Stetson J index* statistic:

$$J_{Stet}(\{x_i\}) = \frac{\sum_i^N g_i \, \mathrm{sgn}(P_i)\sqrt{|P_i|}}{\sum_i^N g_i}, \tag{4.7}$$

where $\mathrm{sgn}(x)$ is the sign function that extracts the sign from the real number x, and g_i is another set of weights defined as:

$$g_i = \left[1 + \left(\frac{|\delta_i|}{2}\right)^2\right]^{-1}. \tag{4.8}$$

The largest values in the distribution formed by the J_{Stet} index are the most probable variable stars (Stetson, 1996). On the other hand, for Gaussian distributed observations, m_i should approach a value around zero.

To complement the quantification of time series $\{x_i\}$ using the J_{Stet}, another statistics developed and proposed using the already introduced relative error parameter δ_i, is the "Stetson K index" K_{Stet}:

$$K_{Stet}(\{x_i\}) = \frac{N \, \sum_i^N |\delta_i|}{\sqrt{\sum_i^N \, \delta_i^2}}, \tag{4.9}$$

which is renowned to be a good indicator of the kurtosis in a time series. This measurement of the kurtosis is also known as the "peakedness" and possess well-defined properties in the case of a Gaussian distributed time series, where $K_{Stet} \to \sqrt{2/\pi}$.

The I_{ws}, J_{Stet}, and K_{Stet} variability indices have been broadly used in the astronomy and astrophysics community to put a label of "variable" to a particular star in this galaxy.

Their mathematical definition and design demonstrated that it is possible to obtain information from the variation of their brightness and realize that the physical processes' timescales are not as eternal (in human lifetimes) as they usually discussed three decades away.

4.2.2 The Von Neumann η index

The η index (von Neumann, 1941a) is a statistic defined as the squared addition of successive differences between adjacent observations over the variance inside the time-serie σ^2_{ts}. The η index is defined by the following expression:

$$\eta(\{x_i\}) = \frac{1}{(N-1)\sigma^2_{ts}} \sum_{i=1}^{N-1} (m_{i+1} - m_i)^2, \tag{4.10}$$

where m_i and m_{i+1} are successive magnitude measurements, N the available number of epochs. The ratio between the square successive magnitudes differences over the variance quantifies the smoothness of a time series. Smooth-distributed time series are expected to have values η approaches to 0.

The properties of the distribution of the η index are well known for a stationary Gaussian distribution. Nevertheless, these properties cannot be directly assumed for unequal sampled time series.

Even so, this statistic has been pointed to as an effective variability indicator in the literature (e.g. (Shin et al., 2009; Sokolovsky et al., 2017)). In (Sokolovsky et al., 2017) was proposed to use the ratio $1/\eta$ as a variability indicator, selecting the larger values from this quantity.

In practice, the cut-off value for selecting variable objects needs to be determined directly from data. For the particular case of VVV time-series, the η index would have value $\eta \sim 2$ for uncorrelated, low-amplitude sources, and extreme values for series with long time variability trends.

4.2.3 Total amplitude ΔK_s index

The total amplitude ΔK_s is a simple expression to quantify the degree or range of variability of an astronomical source. The definition of the total amplitude ΔK_s is defined as the largest difference between the photometric observation, and is described by the following expression:

$$\Delta K_s(\{x_i\}) = \min(m_i) - \max(m_i), \tag{4.11}$$

where the max(x) function returns the maximum value of the time series, hence the faintest observation of the source in question. Similarly, the brightest measurement in the time series is returned using the min(x) function.

Despite the simplicity of its definition, the ΔK_s its been used as a discriminant to identify significant variable sources. For example, in Contreras Peña et al. (2017b) were identified 816 high-amplitude NIR variable sources with $\Delta K_s > 1$ mag, and characterized as likely PMS sources.

The analysis of the distribution of ΔK_s is a valuable and powerful tool to identify likely YSOs amongst irregular and periodic variables stars projected against SFRs.

This "Amplitude index" also performs satisfactorily in identifying sources with a large amplitude like Eclipsing Binary (Algol type, for example) systems and likely pulsating asymptotic giant branch (AGB) stars as Miras and semi-regulars sources, which are especially bright in NIR. These sources have periods longer than P $\geq$ 100 days and are frequently grouped under the name "long period variables" (LPV).

4.2.4 σ/μ ratio index

This statistic quantify the ratio between the mean magnitude $\overline{m}$ and the standard deviation σ_{ts} over time:

$$\begin{aligned} \frac{\sigma}{\mu}(\{x_i\}) &= \frac{\sigma_{ts}}{\overline{m}_w} \\ &= \left(\sum_i^N \frac{(m_i - \overline{m})^2}{(N-1)}\right)^{1/2} \left(\frac{N\sum_i^N w_i}{\sum_i^N w_i m_i}\right), \end{aligned} \tag{4.12}$$

Time series with prominent variability will have significant values in the σ/μ ratio distribution.

The ratio is also known as "coefficient of variation" or "Relative standard deviation" (if you scale the statistic by 100). It is most frequently used in other

math-related fields to understand the prominent variability of a parameter concerning the mean of the population average μ.

4.2.5 Classical χ^2 test index

It was introduced in Rebull et al. (2014a) as a variability indicator to complement the Stetson indices.

This expression is based on the definition on the Pearson χ^2 distribution, or the so-called minimization parameter or goodness of the fit. The χ^2 test is given by:

$$\chi^2(\{x_i\}) = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(m_i - \overline{m}_w)^2}{\sigma_i^2}. \tag{4.13}$$

Of course, this definition measures the sum of their squares differences divided by their uncertainty. The value σ_i is the photometric uncertainty and $\overline{m}_w$ is the weighted average magnitude 4.5.

The χ^2 test is also a comparison between the photometric errors comparable to the signal. This fundamental characteristic makes this statistic sensitive for uncorrelated variable sources. Also, their significant values are reasonably well correlated with stetson indices (Rebull et al., 2014a).

4.2.6 Period determination

One of the most critical and valuable parameters to obtain for the variable sources is the period P of variation (if any). The period immediately gives hints of the timescale variations of the source and reduces the options about the origin of the variability. Of course, there are a variety of physical conditions or systems with particular geometric patterns that can produce periodic or quasi-periodic signals inside astronomical data.

Many efforts have developed methods to identify prominent and significant periods hidden inside a time series. One of the most successful methods for finding a reasonable, reliable period is based on Fourier analysis, under

assumption that the linear combination of trigonometric functions can fully describe a continuous temporal signal $f(t)$ defined as:

$$f(t) = \int_{-\infty}^{+\infty} F(\omega) \exp^{-i\omega t} dt, \tag{4.14}$$

where i is the imaginary quantity, and $F(\omega)$ represents a frequency-domain function that contains the phased amplitudes of the sinusoidal components of ω. The expression is commonly called the "Fourier Transform" and leads to one of the widely used tools in astronomy due to the effectiveness of finding periods in uneven sampled time series: The Periodogram. This set of tools can be used and combined to detect prominent periodicity behaviors inside a set of any irregularly and uneven sampled periodic data and has many applications in other areas of science.

On the other hand, there are also non-parametric methods able to find periodicity in time series with fewer observations. These methods often are based on the absence of theoretical models that explains the signals inside a time series and apply a large spectrum of statistical approaches.

We applied both methods to identify the period P in every generated time series in the considered FoV. To summarize, the two different techniques are described below:

1. The Generalized Lomb-Scargle Periodogram (GLS, (Zechmeister & Kürster, 2009)): A least-squares spectral analysis method based on the classical Lomb-Scargle Periodogram (Scargle, 1982). In particular, we used the `astroML`[1] implementation (Vanderplas et al., 2012).

2. The Informatic Potential Metric Q_m (IP metric, (Huijse et al., 2011)): A non-parametric discriminant designed to identify a fundamental period inside a time series, using the properties of information theory, based on C. Shannon work (1949) and generalized by the Rényi entropy (reference), over the realizations of a continuous random variable.

[1] http://www.astroml.org/

The statistical significance for each determined period is estimated using the spectral power peaks in GLS and IP metric. The peak significance assumes the uncertainties are described by uncorrelated Gaussian noise. Only sources that present a period with peak-significance greater than 99.9% were considered for analysis and further characterization.

Both methods were implemented to complement each other and to cover an extensive range of periods in our data.

The Generalized Lomb-Scargle periodogram

The Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982) is a well-known method for finding periodicity in spurious or irregularly sampled time series data. It is based on the Fourier analysis, where it is possible to find a sinusoidal function $f(t_i)$ to completely describe the behavior of a periodical signal inside a time series under study. The Periodogram fits the time series measurements with a trigonometric model in the least-squares sense, so it minimizes the quadratic error S:

$$S = \sum_{i=1}^{N} [f(t_i) - (A\cos(\omega t_i - \tau) + B\sin(\omega t_i - \tau))]^2, \tag{4.15}$$

where the index i ($i = 0, 1, ..., N$) denotes different measurements $f(t_i)$ taken at time t_i, $\omega = 2\pi f$ is the angular frequency (is its commonly also to found this expression in terms of the period $\omega = 2\pi/\ P$), A and B are the parameters to minimize and τ is a time delay defined by:

$$\tan(2\omega\tau) = \frac{\sum_{i=1}^{N}\sin(2\omega t_i)}{\sum_{i=1}^{N}\cos(2\omega t_i)}, \tag{4.16}$$

this definition of τ ensures that this pair of sinusoidal functions will be orthogonal at time t_i. The parameters A and B represents the contribution or amplitude for each trigonometric term, and describe the of the best-fitting model at frequency ω is stated by the amplitude $\Delta f(t_i) = \sqrt{A^2 + B^2}$ and the phase $\phi = \arctan\left(\frac{A}{B}\right)$.

The strategy to determine the period P hidden in a time series is to evaluate the quadratic error S (4.15) for a range of values of ω and choose the values of ω which minimized S. In their optimal expression, the Lomb-Scargle periodogram for a time series $\{x_i\}$ as a function of ω is given by the following expression:

$$P_{LS}(\{x_i\},\omega) = \frac{1}{2\sigma^2}\left(\frac{[\sum_i m_i \cos(\omega t_i - \tau)]^2}{\sum_i \cos^2(\omega t_i - \tau)} + \frac{[\sum_i m_i \sin(\omega t_i - \tau)]^2}{\sum_i \sin^2(\omega t_i - \tau)}\right), \quad (4.17)$$

where $P(\omega)$ is the Power spectral density, which represents the distribution of the signal power for each ω, and σ is the uncertainty standard deviation of the data. The set of candidate frequencies ω, are multiples of the minimum frequency $\delta\omega$, usually chosen as $\delta\omega = \frac{1}{\varpi T}$, where $T = (\max(t_i) - \min(t_i))$ is the time data range, and ϖ is the oversampling factor, used to adjust the spectral resolution of the periodogram.

For regularly sampled data, a frequency upper limit would be given by the Nyquist-Shannon sampling Theorem (reference). In the case of the irregular sampled data, a rule to select this limit is described by (Press & Rybicki, 1989), which use a multiple of so-called the Average Nyquist frequency:

$$\langle\omega\rangle_{Ny} = \frac{N}{2T}. \quad (4.18)$$

Later in (Zechmeister & Kürster, 2009) was defined the Generalized Lomb-Scargle periodogram (GLS), which includes a constant parameter C in the trigonometric model on the form $g(t) = A\cos(\omega t - \tau) + B\cos(\omega t - \tau) + C$. The new parameter quantified the floor contribution of the spectral power $P(\{x_i\},\omega)$ and improve the estimation of the fundamental frequencies in the cases of the small data sets, or if the data not extend over the whole duration a cycle(VanderPlas et al., 2012).

The Generalized Lomb Scargle periodogram then is defined by the quantities:

$$P_{GLS}(\{x_i\},\omega) = \frac{\chi_0^2 - \mathrm{MIN}_\theta(\chi^2(\omega))}{\chi_0^2} \quad (4.19)$$

where $\chi_0^2 = \sum_{i=1}^{N} \frac{(m_i - \overline{m}_w)^2}{\sigma_i^2}$ is the best-fit sum of residuals (a similar expression

of quantity 4.13), and $\text{MIN}_\theta(\chi^2(\omega))$ is the minimization of the squared differences between the observational data m_i and the model $g(t_i)$ at time t_i; against the parameter vector ` $= [\text{A}, \text{B}, \text{C}]$ inside g(t), which are given by:

$$\text{MIN}_\theta(\chi^2(\omega)) = \sum_{i=1}^{N} \left(\frac{m_i - \text{MIN}_\theta(g(t_i))}{\sigma_i} \right)^2. \tag{4.20}$$

The definition of expression 4.20 and 4.19 ensures the determination of the frequency Power spectrum that lies in the range $0 < P_{GLS} < 1$.

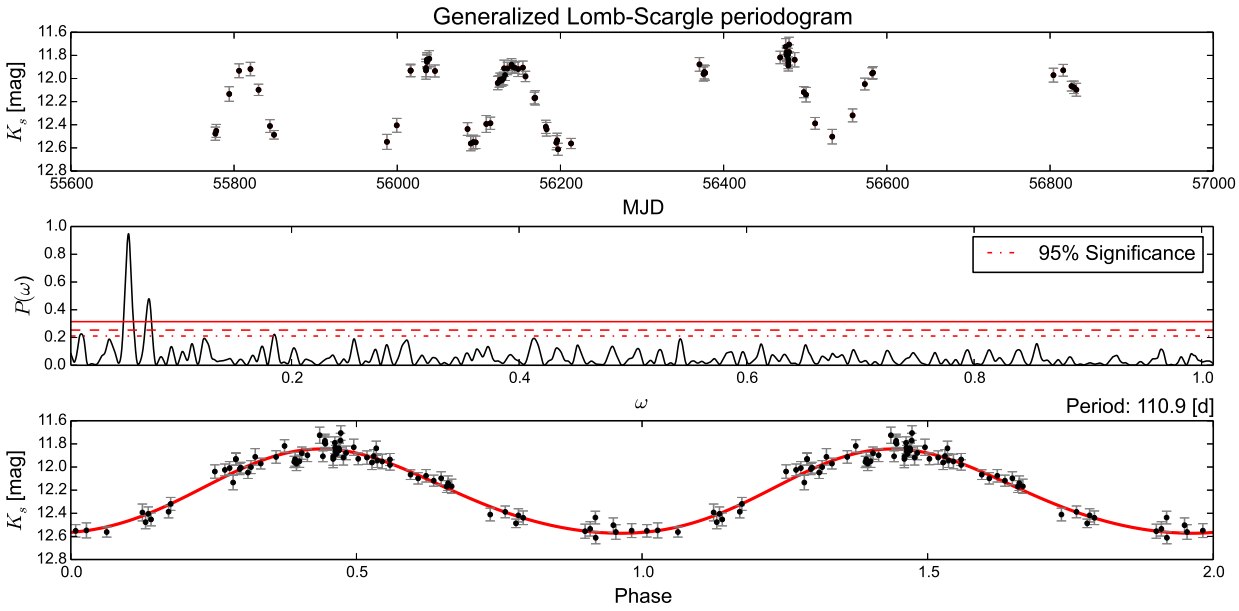


Figure 4.1: Performance of the GLS periodogram. Top plot: Light curve to analyze. Middle plot: Peaks in the spectral Power, where $\omega = 2\pi/P$, the maximum peak have a significance greater than 99.7% (the dashed red line) and a FAP = 0. Below plot: Folded light curve with the best period found (P=110.9 days).

In both cases, the Lomb-Scargle periodogram can detect the period of a time serie using a grid of candidate frequencies ω_{grid} (or periods P_{grid} given the relation $P = 2\pi/\omega$). Therefore, the frequency ω that maximizes the spectral power have an exponential probability of appearing.
We are interested in the statistical significance of a particular ω value that max-

imizes $P(\omega)$ in comparison with the peak of other non-significant frequency values P_0. If we have N_ω number of frequencies for testing, then the probability of a significant value will be Prob($P(\omega) > P_0$) = $\exp^{P_0}$.

On the other hand, the probability of none of the n test frequencies has a significant value $P(\omega) > P_0$ is described by Prob($P(\omega) < P_0$) = $(1 - \exp^{P_0})^n$. Then the quantity Prob($P(\omega) > P_0$) $= 1 - (1 - \exp^{P_0})^n$ represents the siginificance level of the peak $P(\omega)$ that is produced by a test angular frequency ω.

For the after-mentioned properties, we compute the False-Alarm Probability (FAP) based on the definition provided in table 1 addressed in (Zechmeister & Kürster, 2009) also as a discriminator. If we use the probability expression when the spectral power $P(\omega) \in [0, 1]$ the FAP is defined by:

$$\text{FAP} = 1 - \left[1 - (1 - \max(P(\omega)))^{\frac{N-3}{2}}\right]^{N_\omega}, \tag{4.21}$$

where N is the number of observations inside the dataset $\{x_i\}$, and N_ω is the number of tested frequencies, and its proportional to the minimum frequency $\delta\omega$.

Usually, if the power peak has a FAP$<$ 0.0001, the trial frequency ω_{grid} is considered significant. In general, this periodogram analysis is valid for the time series in all ranges of periods.

Information Potential Metric

Claude Shannon introduced the Information theory framework in (Shannon, 1948) as a solution for the noisy channels communication problem. Their primary motivation was to scale the transmission rate according to the noise level inside the selected transmission channel. Shannon discovered that two statistical quantities were enough to reach an optimal encoding message: The mutual information and the entropy.

For an ideal communication channel where it is transmitted a signal S and received (or mapped) signal $\hat{S}$; the mutual information between $\{S, \hat{S}\}$ is maximum when $\hat{S}$ is equivalent to S. In the real world, the noise is everywhere, and it is better to characterize the communication by the channel capacity C, which

is the maxima mutual information between the probability mass function of S and $\hat{S}$ (Principe, 2010).

Assuming we are receiving a message $\hat{S}$, this can be modeled by a random variable X with a set of possible outcomes values $x_X = \{x_1, x_2, ..., x_N\}$, all of them with probabilities $p_k = \{p_1, ..., p_N\}$, when X assumes a certain value. The Shannon entropy (a.k.a. information entropy) then is a measure of uncertainty in a random variable X, and is defined as follows:

$$H(\mathrm{X}) = -\sum_{k}^{N} p(x_k) \log_2 [p(x_k)] \tag{4.22}$$

This expression is the same as physical entropy, but information entropy fundamental comes from the probability mass function, whereas physical entropy is a property of the physical system's state (Principe, 2010).

A mathematical generalization of the Entropy concept is the family of Renyi's entropies of order α (Rényi, 1961). Assuming X is a continuous random variable and $p(x)$ is its probability density function (PDF), the generalized entropy is defined by:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\int p_X^\alpha(x)\, dx \right), \tag{4.23}$$

where the parameter $\alpha \geq 0$. In the limit where the Renyi's order $\alpha \to 1$, the Shannon entropy 4.22 is obtained, so $H_1(X) = -\int p(x) \log_2 [p(x)]$ for $\alpha = 1$.

The case when $\alpha = 2$ is the special interest for their implications. The PDF $p_X(x)$ can be estimate from the Parzen window (Parzen, 1962) using a kernel method, which consists in bring the data to a vectorial space where it is possible find a linear function in order to identify patterns among the inputs. The most widely used kernel is the Gaussian kernel defined by:

$$G_\sigma(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2} \right), \tag{4.24}$$

where x_i and x_j are samples of the random variables X_i and X_j, σ is the kernel size or bandwidth, and $\| \bullet \|$ is an Euclidean norm. Assuming Gaussian kernel, it's possible to obtain a PDF of the form:

$$p_X(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \tag{4.25}$$

and if we use the PDF 4.25 in the quadratic Renyi's entropy ($\alpha = 2$) we can obtain:

$$\begin{aligned} H_2(X) &= -\log\left(\int_{-\infty}^{+\infty} p_X^2(x)\,dx\right) \\ &= -\log\left(\frac{1}{2\pi}\frac{1}{(N\sigma)^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)\right), \end{aligned} \tag{4.26}$$

where the argument within the logarithm is called the quadratic Information Potential (IP):

$$IP_X(\{x_n\}) = \frac{1}{2\pi}\frac{1}{(N\sigma)^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \tag{4.27}$$

and it's a direct entropy estimation of the data, where x_n are the number of measures of the continuous random variable X.

In (Huijse et al., 2011) was defined the IP metric Q_m, which is a discriminant designed to identify the fundamental period among a the possibles candidates P_t. For a signal composed by x_n measures, taken in times t_n $(n = 1, 2, ..., N)$ we used a trial period P_t to transform it into the phase space using the equation (1.2). The folded light curve is segments in a H number of boxes and the IP is computed for every box h. The IP metric $Q_m(P_t)$ is the sum of the squared differences between the information potential of each bin and the global IP:

$$Q_m(P_t) = \frac{1}{H}\sum_{h=1}^{H}[IP_X(\{x_n\}) - IP_X(\{x_n\}_{n\in h})]^2, \tag{4.28}$$

as it mentioned above, when a light curve is folded using it's real period (or a multiple of it) the result in the phase space will be a ordered and well defined signal, what it's translated to a small entropy in each box H. On the other hand, if the light curve is folded with a wrong period, the result will be noise

in the phase space and the entropy in every box will be comparable to the global entropy. Hence, it's possible to find a significant period maximizing the IP metric Q_m.

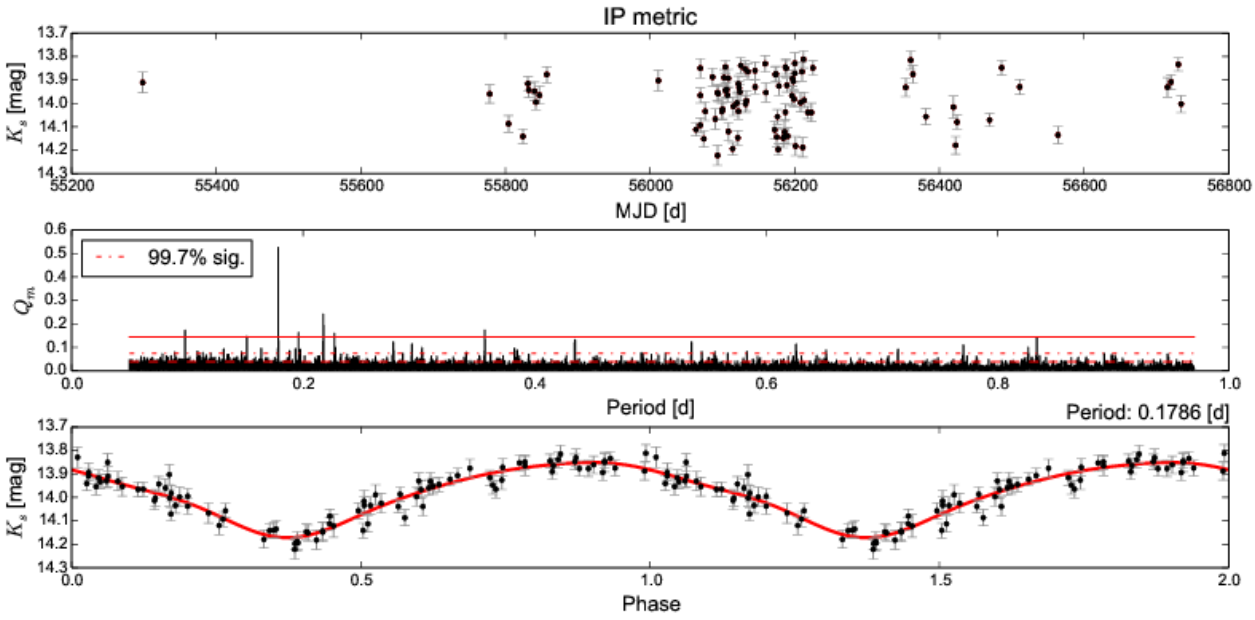


Figure 4.2: Performance demonstration of the IP metric. Top plot: Time serie to analyze. Middle plot: Information potential estimated in the period distribution boxes. The maximum peak have a significance greater than 99.9% (the solid red line). Also it is possible to see more indicative peaks, which are spurious periods and discarded using aliasing correction. Below plot: The light curve with the best period P=0.1786, a very unusual period, but was the period who maximizes the order inside the phase space.

4.2.7 Fourier analysis

Once the period of a particular source is determined, its time series is transformed using the epoch folding technique 1.2. Then their light curve is analyzed and described using a Fourier series. This Fourier fit is constructed by the period P, the average K_s magnitude $\langle m \rangle$, and for the individual amplitudes

A_k and phases ϕ_k for each k harmonics. The Fourier series $f(t)$ at a moment t is defined by:

$$f(t) = \langle m \rangle + \sum_{k=1}^{N} A_k \sin\left(\frac{2\pi k t}{P} + \phi_k\right). \tag{4.29}$$

The variable t is real and defined in the so-called phase space, where $0 \leq t \leq 1$. The performance and goodness of the Fourier fit are quantified using the distribution of the residuals ϵ_i, assuming that it can be modeled using a Gaussian distribution.

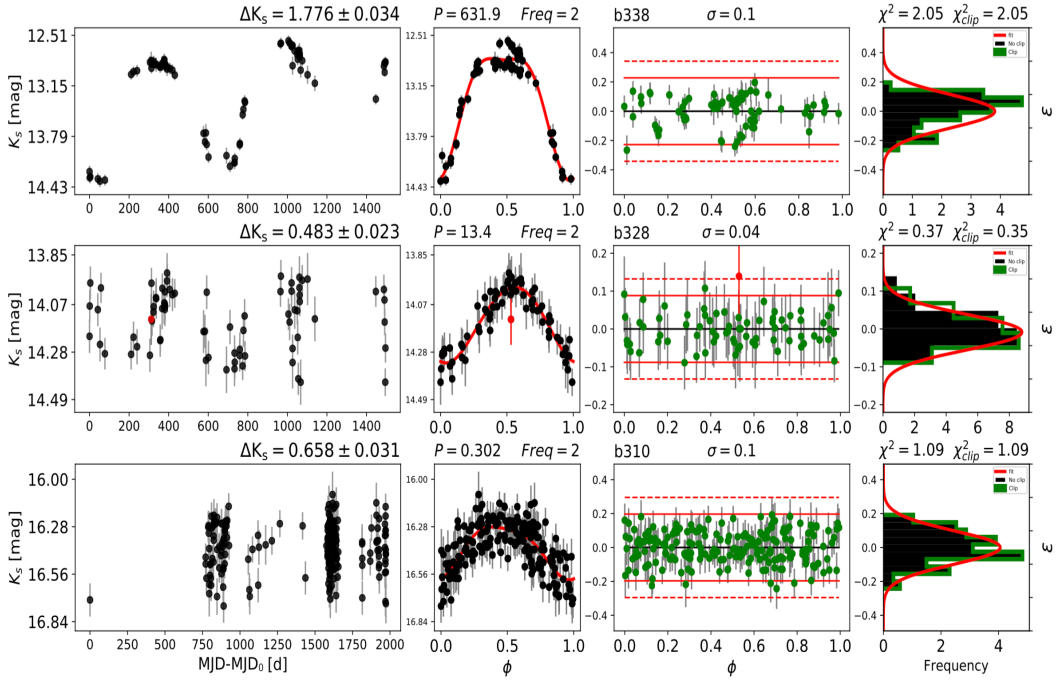


Figure 4.3: Representation of the procedure for the selection algorithm. The raw time-serie is folded using the significative period and the light curve is created. Next a Fourier serie is fitted, and their residuals ϵ are analysed before and after the 3σ clipping process. Red dots in the panels represent the outliers from the fitted Fourier serie.

For simplicity and given the broad range of identified periods, the harmonic number chosen to fit the Fourier series is $N = 2$. This harmonic number selection could not be the best selection for sources with complex taxonomies in their light curve, such as binary systems. For this reason, a 3σ-clipping process was implemented, using the normal distribution model from the residual

distribution ϵ_i. Red dots on the panels of figure 4.3 displays photometric measurements considered as outliers for the Fourier fit.

The reduced $\chi^2_{reduced}$ statistic was used to quantify the performance of the process, using the whole ϵ_i distribution (χ^2) and the residuals without the considered outliers by the Fourier fit (χ^2_{clip}). The complete analysis process for the time-series and light curves is illustrated in figure 4.3.

4.3 Near-infrared Color-color diagrams

For further characterization of the identified variable star sources, the $(J - H)$ vs. $(H - K_s)$ color-color diagram was constructed using the J- and H-band photometry. As it was mentioned in section 2, we used the `Dophot` software to perform PSF photometry of the first observation (2010) of the J- and H-band.

Due to the VVV observational strategy, both J and H images were taken in the same night, with differences ranging from only a few minutes to not more than a few hours. The uncertainties of the photometric measurements are in the order of $\sim$ 0.25 mag at $J \sim$ 19-19.5 mag, and $\sim$ 0.18 mag at $H \sim$ 18.5-19 mag, attributable to a uncertainties of the photometric measurements.

Nevertheless, instead of performing the photometry directly from tile images, we took advantage of the several catalogs that have been made using the VVV survey, such as the VIVA catalog Ferreira Lopes et al. (2020), punctually the photometric measurements to construct the CCM diagrams. It should be noted that the sources do not have significant discrepancies between the comparisons $(J_{Dophot} - J_{VIVA})$ and $(H_{Dophot} - H_{VIVA})$, performing a Gaussian distribution centered on zero magnitude. To achieve the cross-match between catalogs, we used a tolerance of 0.4 arcseconds in all the selected regions covered by the FoV.

It should be noted that the sources do not have significant discrepancies between their magnitudes in the J and H bands, performing a Gaussian distribution centered on zero between their comparison. To achieve the cross-match between catalogs, we used a tolerance of 0.4 arcseconds in all the selected regions covered by the FoV.

The sources of interest are analyzed concerning the locus followed by main-sequence stars, giant stars, and classical T-Tauri stars defined in Bessell & Brett (1988) and Meyer et al. (1997). The reddening vector was characterized using the reddening law of (Rieke & Lebofsky, 1985) first, then they were updated to (Nishiyama et al., 2006). They consider the extinction ratios as follows: $A_J = (0.188 \pm 0.014)A_V$, $A_H = (0.108 \pm 0.008)A_V$ and $A_{K_s} = (0.062 \pm 0,005)A_V$.

We divide the $(J - H, H - K_S)$ plane following the method of Sugitani et al. (2002), and Ojha et al. (2004). Three regions can be distinguished as follows:

1. **F-region**: We define the region between the reddening vectors of main-sequence stars and giant stars as the F-region. These objects are considered to have at most a small NIR excess, likely corresponding to Class III or Class II YSOs. Some of these could be foreground or background sources.

2. **T-region**: The T-region is defined to be between the projected reddening vector from the end of the T-Tauri locus (Meyer et al., 1997) and the main-sequence dwarf reddening vector. Sources located here are considered to have a substantial NIR excess and are likely to be Class II objects, also known as T-Tauri stars. Some Herbig Ae/Be sources with slight NIR excess are also considered to lie in the T-region.

3. **P-region**: The region to the right of the reddening vector extending from the end of the T-Tauri locus is called the P-region. The projected sources in this region are more likely Class I objects, sources with thermal emission from the circumstellar envelope, and protostar-like or Herbig Ae/Be objects.

The membership in a certain region in CCD will only be considered as a suggestion of the evolutionary stage of a source. The nature of the variability can be periodic or chaotic and it is not clear how this can vary the color.

4.4 Outlier rejection and statistical methods

4.4.1 Histogram Bin-width Optimization

The *Histogram Bin-width Optimization method* is a procedure to obtain the optimal bin size for a given set of experimental data considering the analysis from a time-dependent Poisson process. This method was developed in Shimazaki & Shinomoto (2006)[2] for neurophysiologists to analyze the stochastic rate response of a single neuron in a specific range of time.

If we consider that $\hat{\lambda}_t$ is the estimator from a Poisson process λ_t along the entire period of observation T, one method to evaluate the goodness of the fit will be measuring the mean integrated squared error (MISE):

$$\text{MISE} = \frac{1}{T}\int_0^T E(\hat{\lambda}_t - \lambda_t)^2\, dt, \tag{4.30}$$

where E indicates the expectation over different realization from λ_t. If we consider dividing the range T in N bins of size Δ, we can define the expected height of a bar for t inside the range [0,Δ] as:

$$\theta = \frac{1}{\Delta}\int_0^\Delta \lambda_t\, dt.$$

Then, it is possible to manipulate the left part of equation 4.30 in terms of the expected mean rate $\langle\theta\rangle$ to define a cost function $C_n(\Delta)$ by subtracting this term from the original MISE:

$$\begin{aligned} C_n(\Delta) &= \text{MISE} - \frac{1}{T}\int_0^T (\lambda_t - \langle\theta\rangle)^2\, dt \\ &= \frac{2\bar{k} - v}{(n\Delta)^2}, \end{aligned} \tag{4.31}$$

where $\bar{k} = \frac{1}{N}\sum_{i=1}^N k_i$ and $v = \frac{1}{N}\sum_{i=1}^N (k_i - \bar{k})^2$ are the mean and variance from a given set of data $\{k_i\}$, which describes the heights of each bin in the constructed histogram.The optimal bin size Δ will be chosen when the estimator 4.31 will be minimized for a given temporal dataset.

[2] `https://www.neuralengine.org//res/histogram.html#Theory`

Uses: To obtain the optimal bin-size for histograms which contain temporal data processes, as time-series and timescales of physical processes, such as displayed on plots 2.4, 3.1, 4.3, 5.2 5.3 and 5.10.

4.4.2 Modified Thompson τ technique

The *Modified Thompson τ technique* (Thompson, 1985) is a method to determine and remove outliers given a dataset $\{x_i\}$. For each element the absolute deviation $\delta_i = |K_s^i - \overline{K}_s|$ is calculated, where $\overline{K}_s$ is the mean K_s magnitude. The modified Thompson τ statistic is defined as:

$$\tau = \frac{t_{\alpha/2}(n-1)}{\sqrt{n}\sqrt{n-2+t_{\alpha/2}^2}}, \tag{4.32}$$

where n is the number of data points inside the dataset $\{x_i\}$, and $t_{\alpha/2}$ is the critical value of Student's t-distribution given a confidence parameter α.

For each photometric data point K_s in a time series, the standard deviation of the time series σ_{ts} and the absolute deviation $\delta_i = |K_s^i - \overline{K}_s|$ is calculated, where $\overline{K}_s$ is the mean K_s magnitude. Individual photometric measurements were removed from time series when $\delta_i > \tau\sigma_{ts}$, using a confidence level of 95% (i.e. $\alpha = 0.05$). Figure 4.4 shows the performance of this method acting on a time series.

Uses: To remove outliers in the raw extracted time series, we implement the modified Thompson τ technique, which is based on the definition shown in Thompson (1985).

4.4.3 σ-clipping technique

The *Sigma Clipping process*, also known as the σ-clipping technique, is a rejection technique for removing random noise from a particular dataset, identifying which data points could be considered an "outlier." The σ Factor is a quantity that defines the threshold beyond the values that can be regarded as

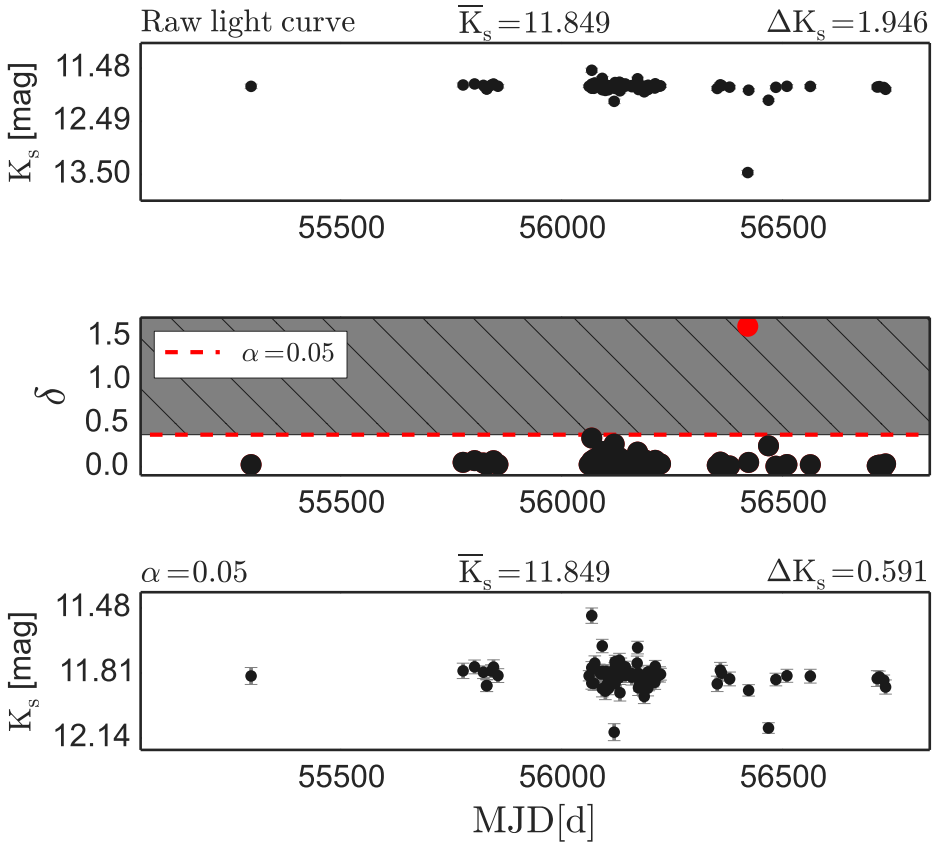


Figure 4.4: Performance of Thompson τ technique acting on a raw time series. Top plot: raw light curve extracted from the photometric process. Middle plot: absolute deviation δ of each measurement as a function of the modified Julian days MJD. The dashed red line indicates the rejection region (gray region) using $\alpha = 0.05$. The red point that falls in this region is removed from the sample. Lower plot: modified time series, which will be used in the analysis.

an outlier. As is said in the GNU Astronomy Utilities [3], the σ-clipping is completely defined as the following iteration process:

1. Calculate the standard deviation (x_σ) and median ($\hat{x}$) of a given dataset $\{x_i\}$.

2. Remove all points that are smaller or larger than $\hat{x} \pm \sigma x_\sigma$.

3. Repeat the process until the convergence criteria is fulfill.

Uses: This technique was used to improve the uncertainty of statistics measured directly from the data. In the linear fit for the calibration of the photometry using the CASU database displayed on 2.3, Also to reject outliers in periodic stars when the Fourier series is used in figure 4.3.

4.4.4 Bootstrap resampling

Inside the non-parametric statistics, the *Bootstrap resampling* technique is a useful statistical method to assess the bias in a specific estimator $\alpha(\{x_n\})$ (i.e. feature, variability index) which is computed from a given dataset $\{x_n\}$, with n the total number of observations within the dataset of a particular source in the sky.

Bootstrap simulates a random sampling process to estimate the uncertainty σ_α over the statistic of interest α, approximating their distribution using only the initial sample $\{x_n\}$.

It is necessary to decide the size of the sample B from each sub-sample X^* and the number of iterations N_{iter} to generate a confidence interval and estimate the bias. These parameters B and N_{iter} will be assessed with the generated VVV time series.

Therefore, the bootstrapping method generate N_{iter} independent subsamples X^* with replacement, of size B. The distribution of these calculated parameters $\hat{\alpha}(X^*)$ maps the uncertainty in the desired statistic from the original sample $\alpha(\{x_n\})$. According to the Law of Large Numbers, the sample

[3] `https://www.gnu.org/software/gnuastro/manual/html_node/Sigma\protect\discretionary{\char\hyphenchar\font}{}{}clipping.html`

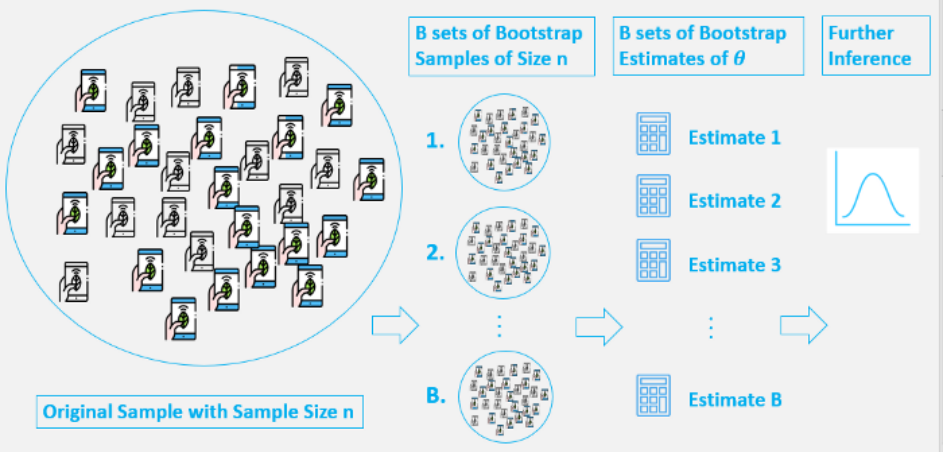


Figure 4.5: An illustration of the Bootstrap estimation and the algorithm to estimate the confidence intervals for the statistic of interest. The image was taken from Lorna Yen website[5]

variance of the statistic $\hat{\alpha}$ should converge to the true variance of the statistic α as N_{iter} becomes larger.

To completely describe the process, algorithm for the Bootstrap procedure follows:

1. Obtain N_{iter} random sub-samples X^* of size B from the original dataset $\{x_n\}$.

2. Compute the desired statistic $\hat{\alpha}$ for each N_{iter} sub-sample.

3. Obtain the mean value $\overline{m}$ and variance value s^2 from the new generated distribution $\hat{\alpha}$, where we defined the mean $\overline{m}_{\hat{\alpha}}$ and variance $s^2_{\hat{\alpha}}$ value as usual:

$$\overline{m}_{\hat{\alpha}} = \frac{1}{N_{iter}} \sum_{i}^{N_{iter}} \hat{\alpha}_i \tag{4.33}$$

The variance will be:

$$s^2_{\hat{\alpha}} = \frac{1}{N_{iter} - 1} \left\{ \sum_{i}^{N_{iter}} (\hat{\alpha}_i)^2 - \left(\frac{1}{N_{iter}} \sum_{i}^{N_{iter}} \hat{\alpha}_i \right) \right\}^2 . \tag{4.34}$$

However, it is important to note that the best estimator for the statistic of interest α would be measured for the complete set of data $\alpha(\{x_i\})$.

Uses: To obtain the empirical bootstrap estimate to estimate the corresponding confidence interval. For parameters or to ascertain the error on a parameter estimated from time-series data points, as the mean K_s-band $\overline{K}_s$ or the amplitude ΔK_s as examples.

4.5 Hardware and performance

Considering the size of each VVV tile image is about 800 Megabytes per each `FITS` image, it is reasonable to understand the challenging task of obtaining the PSF photometry from their analysis, particularly in regions close to the most crowded areas of the galaxy.

To analyze such data volume described in table 3.1, it was necessary to organize and prepare the data flow in terms of data structure, analysis, and the storage of the information related to the research. Several routines were mainly written in `Python` (version 2.7 and 3), and `Bash script` were developed and assembled to fulfill this task.

Just a few more words about the process described briefly in section 2.1: First, the VVV tile of interest was identified, then all the available files (`FITS` files and aperture catalogs) of the requested field were obtained using the `wget` method. All obtained images were compressed to a value close to a quarter of their actual size. The `imcopy.c` script available on CASU data was used to access the real-size tile images.

To prepare the data after the application of `Dophot` software, the header of the images was consulted to obtain information such as the bandpass, information about the weather, and the specific size. The sky background was directly estimated in each image, dividing the image recursively into small squares where the background value was assessed and averaged. In this case, the uncertainty of the value of the sky was purposely ignored.

In parallel, the time series of all available sources were created with the aperture photometry catalogs to obtain the set of standard stars. At this point,

the aperture catalogs were no longer helpful and were deleted. The next and final step of the first stage of the process is the calibration of the instrumental photometry and the creation of the multi-epoch master catalog. This process takes about three days for a tile on the disk (between 50-60 epochs), around six days for a tile inside the bulge (between 80-90 epochs), more than a week for a high-cadence tile inside the Baadee's window (between 280-310 epochs).

With several gigabytes in size, the multi-epoch catalog was transferred to the computing cluster of the University of Valparaíso (UV), where the second part of the process was done using between 30 and 45 cores in parallel.

First, the individual file with the time series for each detected source was created and characterized, measuring the complete set of features described in section 4.2. Finally, the two periodograms (GLS and IP metric) were applied for every extracted stellar source. These last processes were costly in terms of computational calculus, so completing the second part took a few days in the case of tiles projected in disk, and even more than a month, for high-cadence tiles.

The hardware used and their communication is shown in the image 4.6. There were four personal computers configured with Ubuntu operating system used for the first part of the process, the PSF photometry. These were: `Dell_Nuevo`, `Alumno`, `Mac_ubuntu`, and `MacBaren`. These four personal computers were linked through the `Dropbox` software to share information and leave notifications about the status of the photometric process. In addition, the computers described also had access to the UV computing cluster, where the multi-epoch catalogs were ready for analysis.

Finally, several external hard drives were also used to save and back up the information of the processes that had already been completed.

In this work, we analyzed 1411 tiles images, with more than 1,12 Terabytes of information analyzed and more than 24.6 million time series analyzed in all the considered FoV.

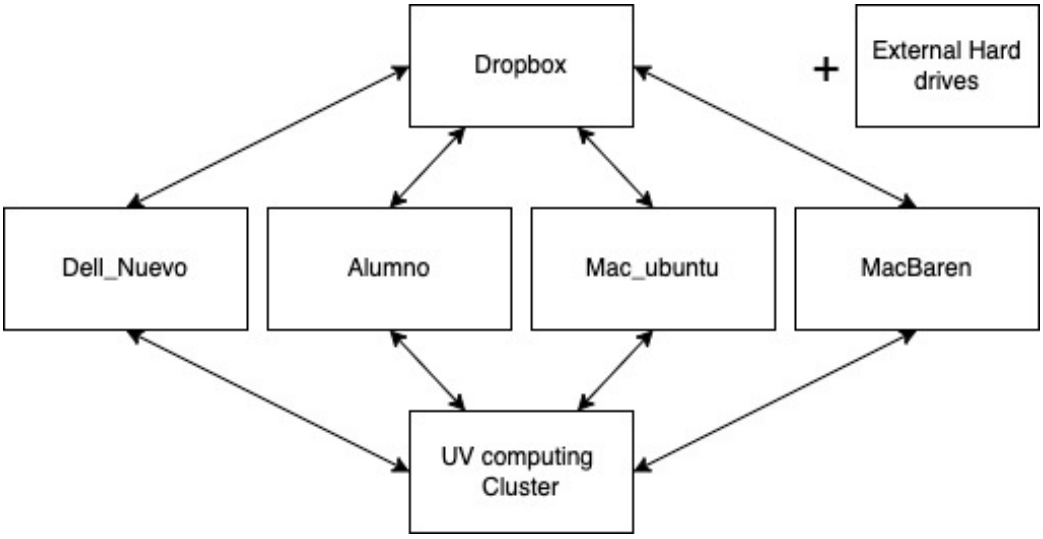


Figure 4.6: Schematic view of the automated process developed and used in this study to categorize the sources in this manuscript.

CHAPTER 5

Results

Variability is a powerful technique to identify the unstable nature of sidereal objects across the universe. The main reason for identifying YSOs through variability is that they dominate the IR variability at high amplitudes within SFRs. Highly variable sources are related to accretion and eruptive episodes in class I and II YSOs, given they are surrounded by shells of interstellar material.

Still, the variability scenario can also be the manifestations of differential extinction, material occultation, and hot/cold spots on the surface of the young star. The different nature of these scenarios will produce expected features or behavior in their observed time series. Likely YSOs can be identified using this characteristic and quantified behavior in time.

We expect that we can find overdensities of variable stellar objects in areas with other signs of stellar formation, such as ionized hydrogen regions, Masers, open clusters, accumulation of interstellar material.

Variable sources of any possible nature could appear and are expected: Red giant stars, novas, microlensing, and even background galaxies as quasars can also display large IR variability. Also, pulsating and periodic stars, in general, have low amplitudes in the NIR and fast modulations in their brightness. Therefore the shape in the phase space defined on section 1.2 may vary.

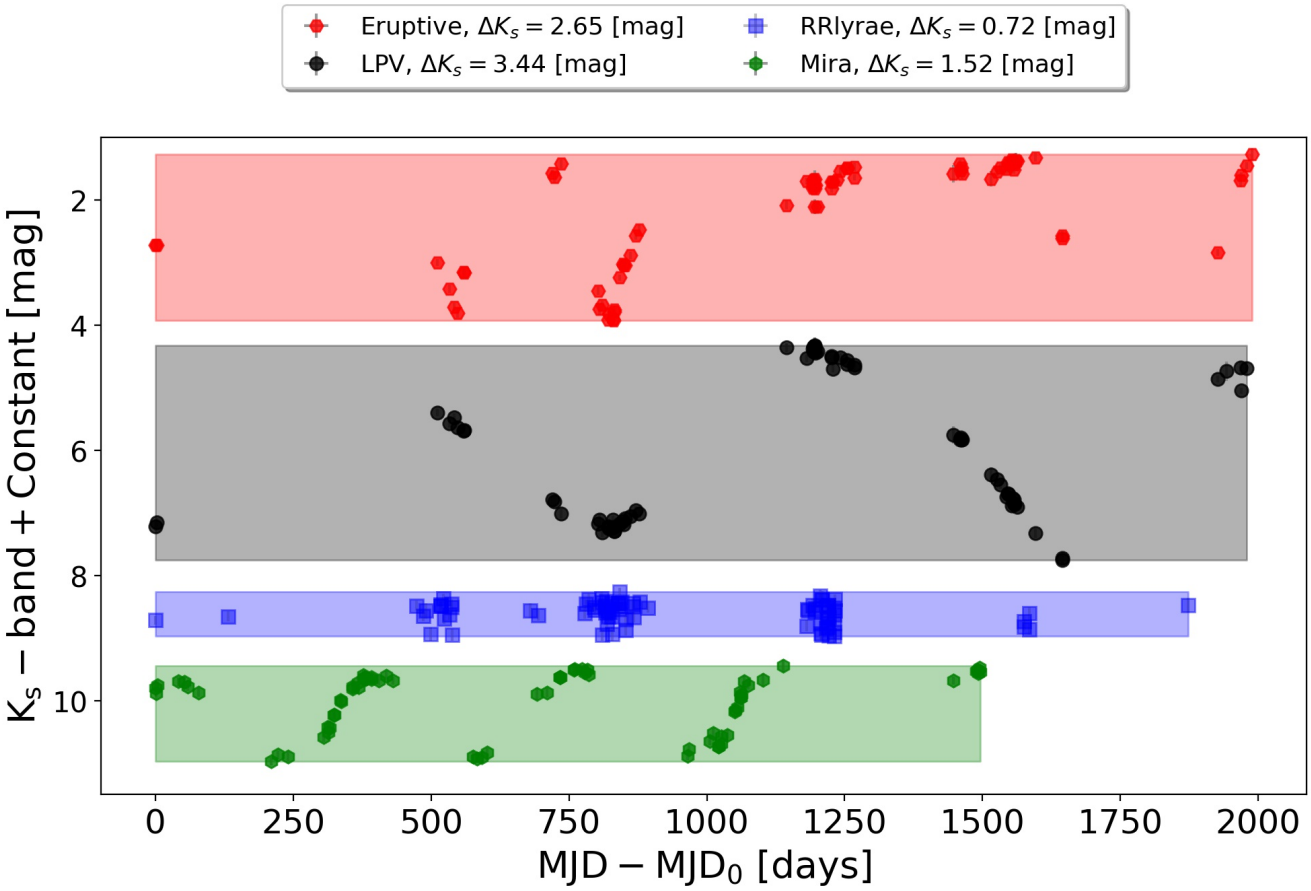


Figure 5.1: Comparison of the variety of morphology and photometric amplitudes of the selected variable stars inside the considered FoV. These selected stars are sorted in amplitude from the largest LPVs tops, then an Irregular variable source (red dots), a Mira-like periodic star, and the humble RRlyrae in blue. This is a direct comparison of the amplitudes of different physical systems.

Figure 5.1 shows the different variability profiles for different scenarios of stellar variability. It is a direct comparison of the amplitudes scales about their varying nature. The MJD has been normalized to the first observation inside their datasets. At this scale, the error bars measurements are almost negligible. The height of each colored rectangle represents the Ks-Band amplitude ΔK_s.

The top time series (red dots) displays the increasing and decreasing nature of the **Irregular variable source** SFR4_116.

The LPV variable SFR3_pv19 is displayed next. These evolved stars are famous for their particularly large periods and amplitudes at these wavelengths, and it is possible to visualize their periodical behavior along the 2000 days baseline. As can be seen, only the LPVs amplitudes are comparable with amplitudes of the irregular variable star objects.

In the case of the Mira-like source SFR3_pv743 (green rectangle) is also possible to visualize their period, but this is not the case for the RRLyrae star as the variable named SFR4_pv4 identified as an RRLyrae star (blue rectangle). With low NIR amplitudes and short periods ($0.2 < \text{Period} < 2$ days). SFR4_pv4 time-serie looks like noise in comparison to the other variable sources.

Both high-amplitude sources SFR3_pv743 and SFR3_pv19 were identified on VVV tile b324 and share the same observations logs. The RRLyrae stars were placed on VVV tile b341 and the Irregular variable star SFR4_116 on b338. The sampling differences between tiles are appreciable in their unevenly sampled time series.

The amplitude ΔK_s described in section 4.2.3 will play an essential role in the sources selection. It is the main physical parameter that we can detect from a variable YSO.

5.1 High amplitude irregular sources

The searching of YSOs through NIR photometric amplitude is a widely accepted technique (e.g., Borissova et al. (2016); Teixeira et al. (2018); Contreras Peña et al. (2017c); Rebull et al. (2014a); Wolk et al. (2015)). For example, in Contreras Peña et al. (2017a), the total VVV K_s amplitude ΔK_s was used as

a discriminant to identify variable sources with $\Delta K_s > 1$ magnitude. This method was handy to help identify likely YSOs amongst irregular and periodic variables stars projected against SFRs. High-amplitude sources may reveal Accretion-driven outbursts in young cores surrounded by shells of interstellar material. On the other hand, low amplitude sources would inform the intermediate state of the accretion process and relatives with the surrounding material.

For the Amplitude distribution ΔK_s, it was necessary to develop an adaptable and data-dependent analysis, given a large number of detected stars per FoV. The distribution of ΔK_s was characterized using a non-parametric fit to determine the behavior of ΔK_s as a function of K_s in the number of bins determined by section 4.4.1. Then the dispersion in $\sigma_{\Delta K_s}$ in ΔK_s as a function of the average value K_s was measured. Sources with amplitude above $4\sigma_{\Delta K_s}$ were selected for further analysis.

In the case of the η index, their distribution is well known as a stationary Gaussian distribution (von Neumann, 1941b), The η index is helpful for quantifying the independence of successive data points inside a time series. For uncorrelated photometric measurements (i.e., time series with uncorrelated normally distributed measurements), the η index would have a value ~ 2, and extreme values for series with long time variability trends. Given the massive number of recovered sources, it is possible to apply the law of large numbers and assume that the index comes from a Gaussian distribution; therefore, the σ parameter of the fitted distribution was used as a proxy for the standard deviation. Sources more than 3σ below the mean were considered for the method.

Panel left on figure 5.2 taken from (Medina et al., 2018) describes quantification ΔK_s using a non-parametric fit. The figure displays the selection of variable star sources inside d001 and d002 VVV tiles. The parameter $\sigma_{\Delta K_s}$ increases for the selection of faintest sources. The background distribution (in black) is the tile d001 population: more than 500 thousand sources map the variability inside the FoV. In the particular case for d001 and d002 VVV tiles, we selected star with $\Delta K_s > 0.6$ magnitudes. The right panel displays the η distribution in function of K_s divided in bins of 0.5 magnitudes. At the beginning, both tails of the normal distribution were used for select variable sources.

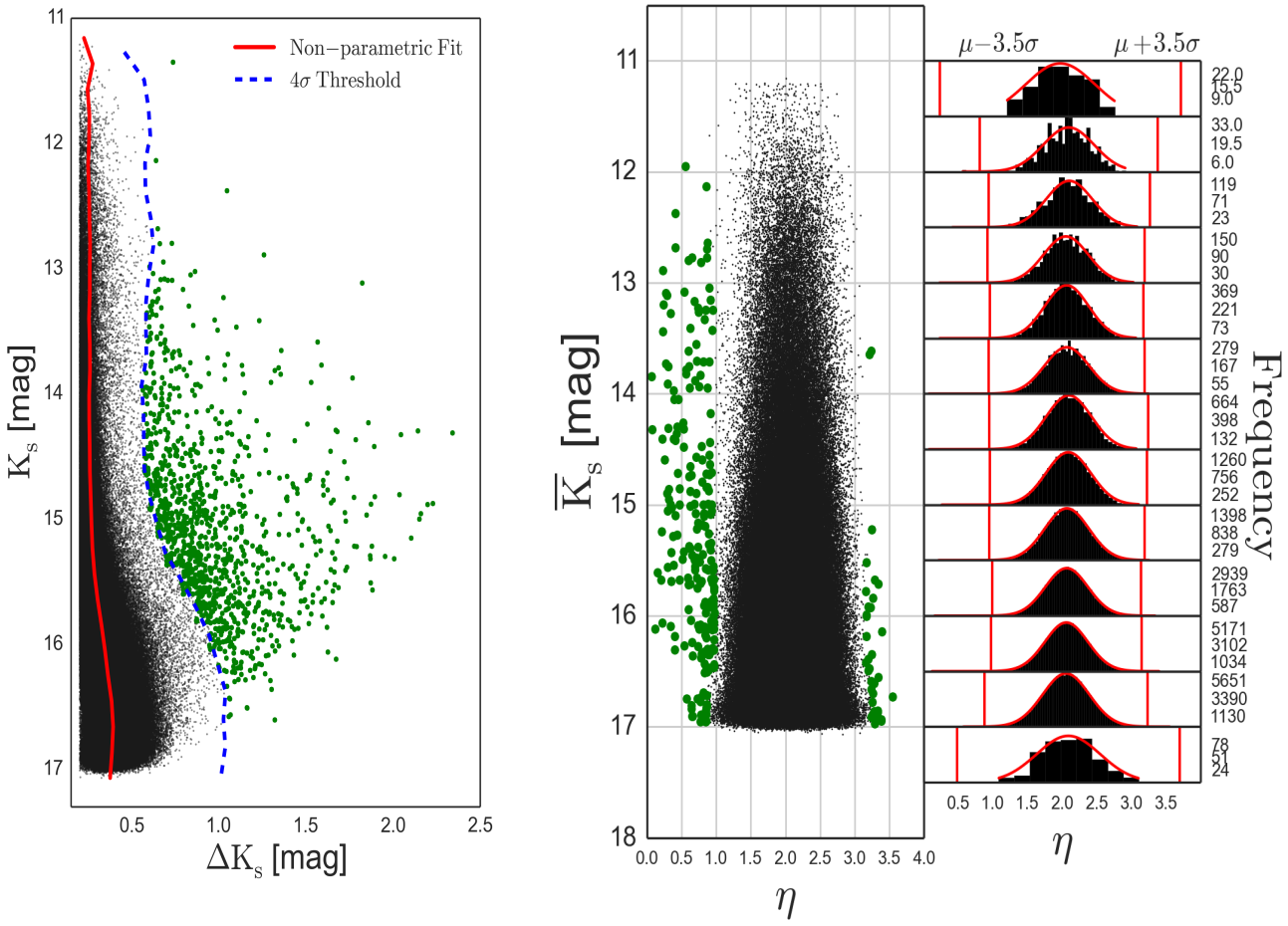


Figure 5.2: Representation of the two selection criteria aiming at identifying irregular variables. The dark points are all stars in the final photometric catalog, and the green filled circles highlight the selected sources by each method. Left panel: Total amplitude ΔK_s selection. The red solid line represent the non-parametric fit, while the dashed blue line displays for the 4σ threshold. Right panels: η index selection. The red lines in the bins of 0.5 mag represent the 3.5σ threshold of the Gaussian fit.

Nevertheless, sources with $\eta > 2$ do not fulfill the amplitude criteria generally; and they were identified as related to saturated objects. Sources selected with $\eta > 2$ were discarded, given the fact that η value tends toward zero indicates a strong variability (Sokolovsky et al., 2017).

In this case in particularly there were selected variable sources with η values < 0.95.

When a variable source has dispersion greater than the above-determined criteria, the source will be considered an **irregular variable star** and their photometric properties will be researched.

As discussed in Introduction, different processes can be responsible for luminosity variations in young stars. One of the main issues in this field of research is determining the origin of those brightness changes. Usually, the variability results from the contribution of various physical mechanisms, depending on the circumstances and environments of each particular source. Sometimes, this contribution can be dominated by one of these mechanisms. Then it is possible to associate the morphology of the changes as a function of time with a specific type of variability. For example, (Carpenter et al., 2001) constructed a simple model of variations in brightness induced by cold spots, assuming a photospheric effective temperature $T_{\rm eff} \sim 4000$ K for stars with age ~ 1 Myr and mass ~ 1 M$_\odot$. According to this model, if the spot coverage in the photosphere is assumed not to be larger than 30%, the amplitude caused by this mechanism can be $\Delta K_s \sim 0.3$ magnitudes.

Figure 5.3 shows the amplitude ΔK_s distribution of the sources selected in the different SFRs as a function of their mean K_s magnitude. In the background as black dots, the amplitude distribution obtained from a tile within the FoV is shown for comparison.

The effects of saturation in the K_s-band are visible in the distribution for sources brighter than $K_s \sim 11.5$ mag and amplitudes ΔK_s larger than 0.55 magnitudes, on average. If a source is identified in this region, it will be carefully verified before its final selection.

The histogram of the ΔK_s distributions for each SFRs is displayed on the right panels. Models such as the cold spot could explain part of the peak of the ΔK_s distribution around 0.3 to 0.6 magnitudes inside the four different distri-

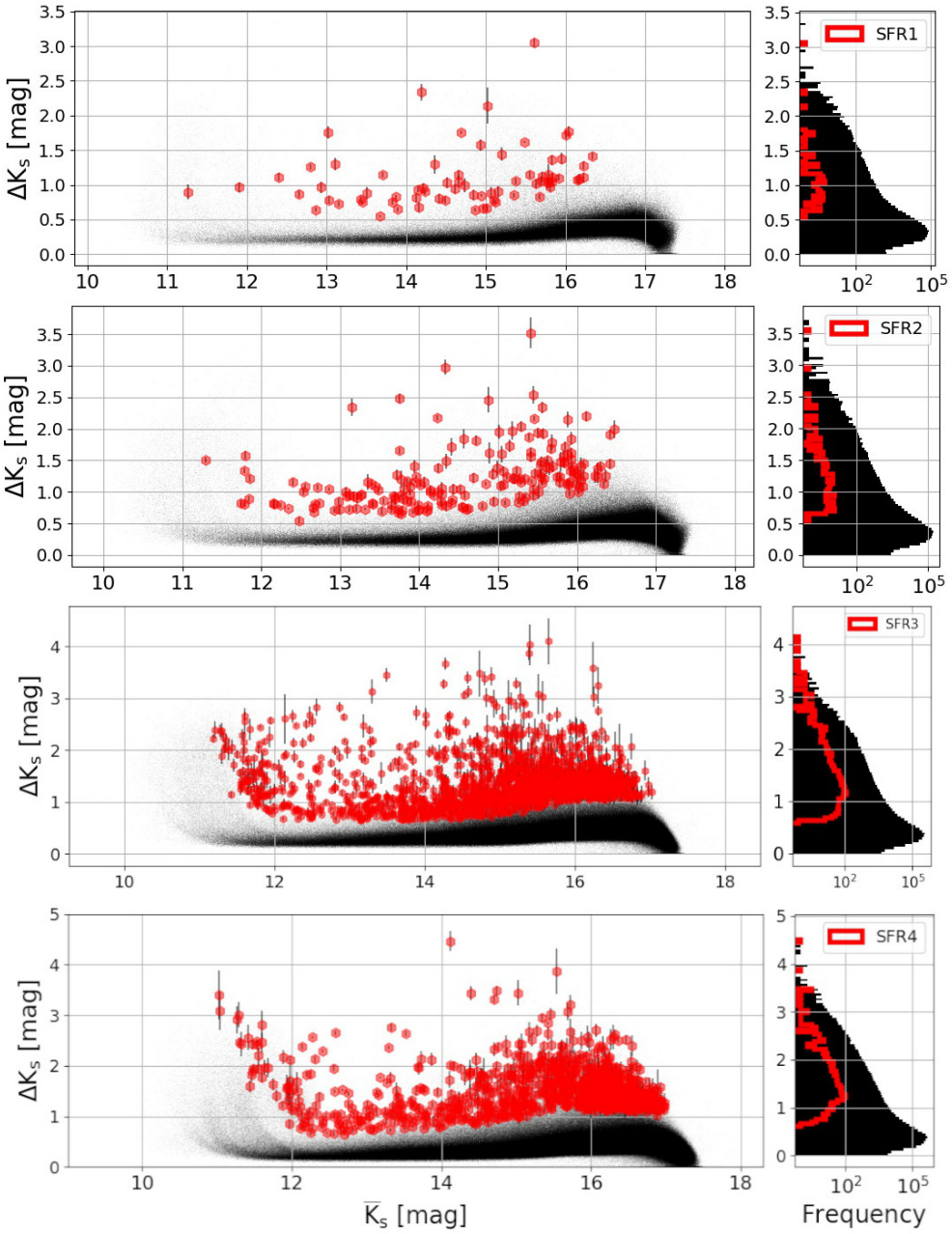


Figure 5.3: Amplitude ΔK_s as a function of the average K_s magnitude of stellar sources that fulfill the amplitude criteria (red hexagons) and the sources with time series (black dots). From top to bottom, the plots of SFR1, SFR2, SFR3, and SFR4 are displayed, respectively. At the right part, the distribution of ΔK_s is shown. The selected sources by the Amplitude criteria are included in the red lines. Also, it is possible to see the increase of the number of sources in each SFRs beyond the limit 10^5 in the amplitude range 0.3-0.5 mag.

butions constructed.

In the case of SFR1 (the top plot) the amplitude interval of the selected sources (red dots) is between $0.56 < \Delta K_s < 3.21$ mag with an average value of ΔK_s =1.1 mag. The same procedure was used in the other SFRs, and similar results were also obtained.

5.2 The time series profiles

Several authors have quantified the morphology of these time series to create a connection with specific processes. Still, this classification depends strongly on the type of data that we are analyzing. As an example, in (Rebull et al., 2014a), (Wolk et al., 2015) and related papers within the YSOVAR project framework defined several types of variable stars using the MIR time series from the Spitzer Space Telescope. They discussed, for example, that rapid time scales suggest disk perturbations, commonly observed in PMS stars. Also, the reported stars classified as dippers are closely related with prolonged dimming effects and are probably caused by dust clouds or perhaps the disk itself. Therefore, is highly probable that high-amplitude variability on timescales from months to decades in pre-main sequence objects is usually associated with episodic accretion events, attributed to disk instabilities typical of the early stages of young stellar objects.

For the VVV survey, NIR irregular variable sources have been split into six categories related mainly to the time scales of the variations and the overall, most common, states: Eruptive, Faders, Dippers, LPV-YSOs, Short Time-scale Variable (STV), and Low Amplitude Eruptive (LAE). The classification aims to highlight the main physical process that controls the variability. This scheme was used in Contreras Peña et al. (2017b), Medina et al. (2018), Medina et al. (2021) and references therein, in an effort to identify variable sources with VVV time series. Of course, some of the sources do not fit in the of categories described above and remain unclassified.
The main attributes for defining each category are described below:

- **Eruptive variables**: Describe high-amplitude ($\Delta K_s > 1$ mag) eruptive

episodes in a stellar source. Accretion-driven magnetic activity can explain some eruptive phenomena, as flares and cold spots outbursts in young cores undergoing the magnetospheric accretion process.

- **Low amplitude eruption variables**: An extension of the Eruptive variables category for sources with display amplitudes lower than one magnitude ($\Delta K_s < 1$ mag). Sources present outbursts with amplitude lower than one mag and duration typically longer than a year.

- **Dippers**: Dippers sources are characterized by long-lasting dimming events followed by a return to average brightness. Dippers are often associated with increased extinction from the surrounding material.

- **Faders**: Sources labeled as Faders can be caused by either a return to a quiescent accreting phase or a long-lasting increase in extinction. Faders show time series that slowly decline over time or a period of constant brightness, followed by a sudden decrease sustained over a year.

- **Short Time-scale Variable**: Sources with Short Time-scale Variability (STV) present rapid increases/decreases of brightness with variable amplitude between individual events. STVs describe sources with eruptions on time scales of hours to days.

- **LPV-YSOs**: The LPV-YSO sources are stars with a well-defined period longer than 100 days and present persistent short timescale scatter in their time series.

Examples of time series of different variability categories taken from SFR2 in the G305 complex are displayed in figure 5.8 and 5.9.

Recent studies (e.g. Guo et al. (2019)) found periodicity in some of the YSOs, but in general the high amplitude variations do not show well-defined periods. More details about categories are summarized in Table 5.1.

The Irregular Variable stars sample inside the SFR1 are displayed in figure 5.5. The most significant part of the sample belongs to tile d001 and follows the cold gas/dust distribution as traced by the W3 ($\lambda_{\text{eff}} = 12m$) band WISE

Class	Description
Dippers	Shows fading events, to then return to their normal magnitude.
Eruptives	Shows sources with outbursts with amplitude > 1 mag and duration longer than a few days and typically at least a year.
Low Amplitude Eruptive	Sources that present outbursts with amplitude lower than 1 mag and duration typically longer than a year
LPV-YSOs	Sources with a measured period, but with short-timescale scatter in the time series.
Short Timescale Variables	Sources with fast and constant scatter in their time series. They also can show brief rises in the magnitude in time scales of weeks.
Faders	Shows a continuous decrease in brightness (t >1 yrs), or a big decrease in its brightness in a source with relatively constant luminosity.

Table 5.1: Characterization of the irregular variables in the categories proposed by Contreras Peña et al. (2017b).

image. The shape of the dots represents their variability categories. The highest over-density is observed at the borders of the star-forming region, where the nebulosity is overwhelming, thus suggesting active star formation.

5.3 Color-color diagrams

The Irregular Variable sources are found in different interstellar environments and one way to roughly estimate their evolutionary stage is by using infrared color-color diagrams.

To analyze the position of our irregular variable sources on the color-color diagram, we have followed the method described in Ojha et al. (2004). Three regions are defined: The 'F' region, which is located between the reddening vectors of giant and dwarf stars. The 'T' region is between the reddening vector of giant stars and CTTs locus, where the Class II YSOs objects and Herbig Ae/Be stars (Hillenbrand et al., 1992) can be identified. In the so-called 'P' region, located below the reddening vector of CTTs, the likely proto-stellar objects are situated. The solid black lines are the intrinsic colors of dwarf and

giants stars from Bessell & Brett (1988), the solid red line is the locus of un-reddened CTTs Meyer et al. (1997), and the dashed red lines are the reddening vectors of the early spectral type dwarfs, giants stars and un-reddened CTTs. The reddening vectors extinction law and assuming a visual extinction A_V =15-25 mag for each evolutionary track are also shown (see figure 5.4).

In the so-called P-region, protostellar sources (i.e., Class I YSOs) are expected to lie redward of the reddening vector of the classical T Tauri (CTT) locus, given they are intrinsically red due to their evolutionary state. Given their IR variable nature, this is consistent with the type of stellar objects we expect to identify. To create the CCD with the identified sources, we obtained the J-band and H-band measurements using the VIVA catalog (Ferreira Lopes et al., 2020) of all measured Irregular variable stars. About the half of the total sample (524 from 1096) do have J-band and H-band measurements in different SFRs, and their location in the CCD was quantified.

The membership histogram for all selected sources is displayed on the left side of the figures 5.4. The percentage of each region relative to the total number of sources shown in the CCD is indicated.

Considering all the SFRs, the T region reaches approximately 34.8% of the samples. The P region is populated by the 23.2% of the total sample, against the 42% of the F sample, expected to be elsewhere in the galaxy.

Initially, a 42% of the sample within the F category may seem excessive for stars that are expected to be intrinsically reddened. Nevertheless, two things must be considered: 1) the sum of the P+T categories reaches more than 58% and 2) more than half of the Selected sources do not have color, which could indicate that their colors are unavailable because they are highly obscured inside IRDCs.

SFR1 and SFR2 have the lowest quantity of selected sources (44 and 104, respectively) with available colors, and their stars are almost homogeneously distributed in each category considered (between 31%-36%). On the other hand, the SFR3 and SFR4 FoVs have a more significant number of sources within the F region, and their CCDs will be discussed later.

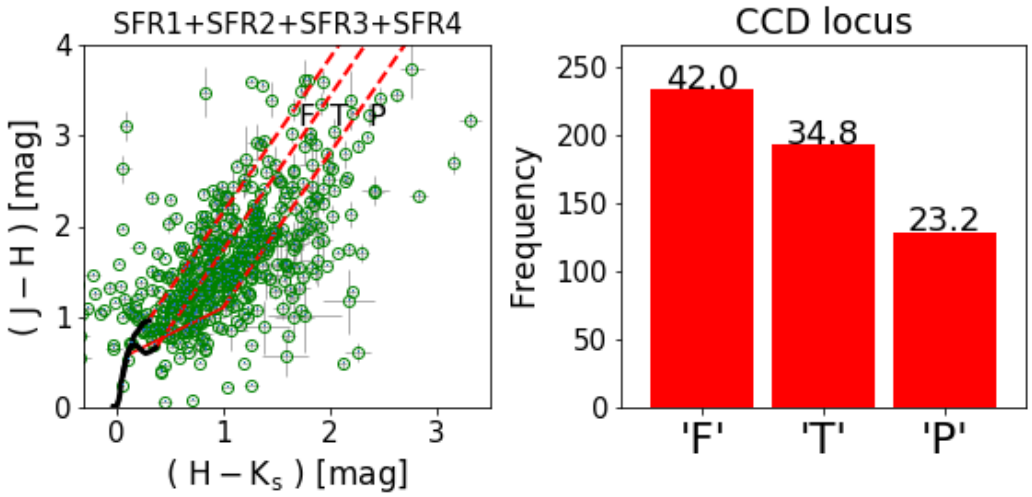


Figure 5.4: Left: the color-color diagram of the 524 selected sources with available colors inside all SFRs. The solid black lines are the intrinsic colors of dwarf and giants stars from Bessell & Brett (1988), the solid red line is the locus of un-reddened CTTs (Meyer et al., 1997), and the dashed red lines are the reddening vectors of the early spectral type dwarfs, giants stars and un-reddened CTTs, assuming a visual extinction $A_V = 15 - 25$ magnitudes. Right: The histogram of the F, T, and P region classification of the sources with colors. At the top of each bar, the percentage of membership in each category is displayed.

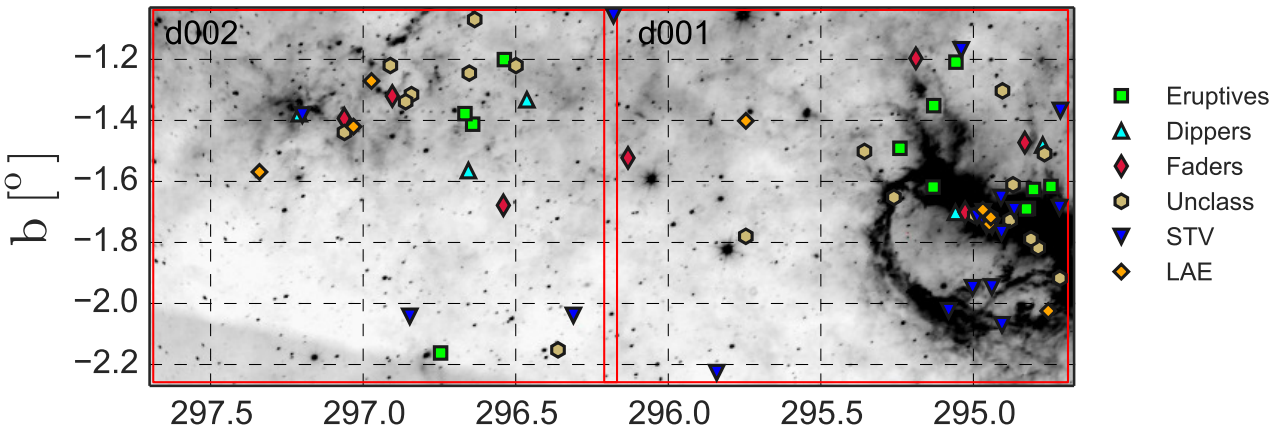


Figure 5.5: Covered region by d001 and d002 tiles from VVV survey, showing the spatial distribution of identified sources. Galactic north is up, galactic east is to the left. The symbols represent different types of variable stars found in the FoV. In background, W3 band WISE image is shown to illustrate the cold gas/dust distribution in the FoV.

5.4 Individual Star-forming regions

5.4.1 SFR1: The surroundings of IC 2944/294

In Medina et al. (2018) we reported the variables stars found in SFR1. A total of 1,308,626 point sources between $10.8 < K_s < 17.2$ mag were detected inside this region. The 898.736 sources had more than 25 epochs and an amplitude greater than 0.2 magnitudes, and their time series were constructed and analyzed.

We identified 70 irregular variable sources that fulfill both the ΔK_s and η criteria: 45 of them belonging to tiles d001 and 25 to d002. Almost two-thirds of them (64.28%) are projected in the d001 tile, following the cold gas/dust distribution as traced by the W3 ($\lambda_{\rm eff} = 12\mu m$) band WISE image (the background of Figure 5.5).

The highest over-density is observed at the borders of the star-forming region, where the nebulosity is overwhelming, suggesting the active star formation status on the field. Some of these sources are shown in Figure 5.6. About 90% of these sources are previously unknown as variable stars.

The top plot of figure 5.3 displays the ΔK_s distribution of the selected sources as a function of the mean K_s magnitude. The selected sources have an average K_s-band magnitude distributed between $11.2 < K_s < 16.4$ mag, and a total amplitude contained between $0.5 < \Delta K_s < 3.2$ mag an average value of $\langle \Delta K_s \rangle = 1.1$ mag. Since the effects of saturation in VVV are substantial in the distributions for sources brighter than $K_s \sim 11.5$ mag and amplitudes ΔK_s larger than 0.65 magnitudes, these sources are carefully verified.

They were classified using the framework described in 5.2. Sources that could not be classified unambiguously into the categories mentioned above are marked as "unclassified". We classified 20 STVs, 12 Eruptives, 5 Dippers, 7 Faders, 8 Low Amplitude Eruptive Variables, and 18 sources remained unclassified.

As an additional filter, if a source presents periodicity, it is removed from the sample and added to the periodic sample. This was the case of two sources (d001-79, d002-103) which present a large amplitude and period, typical signatures of dust-enshrouded AGB stars, invisible in the optical range due to its thick circumstellar envelope, a product of its high mass-loss rate.

Nine open cluster candidates are projected in the field of view of d001 and d002, namely VVV CL005, 007, and 009 Borissova et al. (2011) and La Serena 001, 002, 003, 009, and 015 Barbá et al. (2015). Inside the VVV CL005 cluster radius are the most significant number of irregular variables as probable cluster members (d001-25, 27, 28, 29, and 30). The cluster LS002 has two irregular variables, while clusters LS001 and VVV CL007 have one irregular variable, each other, as probable cluster members. All irregular sources projected close to open clusters in the region are young stellar object candidates.

5.4.2 SFR2: The G305 complex

G 305 is a relatively close-by, luminous giant HII region with active star formation. Two studies were published by us: In Borissova et al. (2019), we used moderately high-resolution spectroscopy obtained with the Apache Point Observatory Galaxy Evolution Experiment-2 (APOGEE-2; Majewski et al. (2017)) spectrograph to study the massive stellar content in the area. We investigated

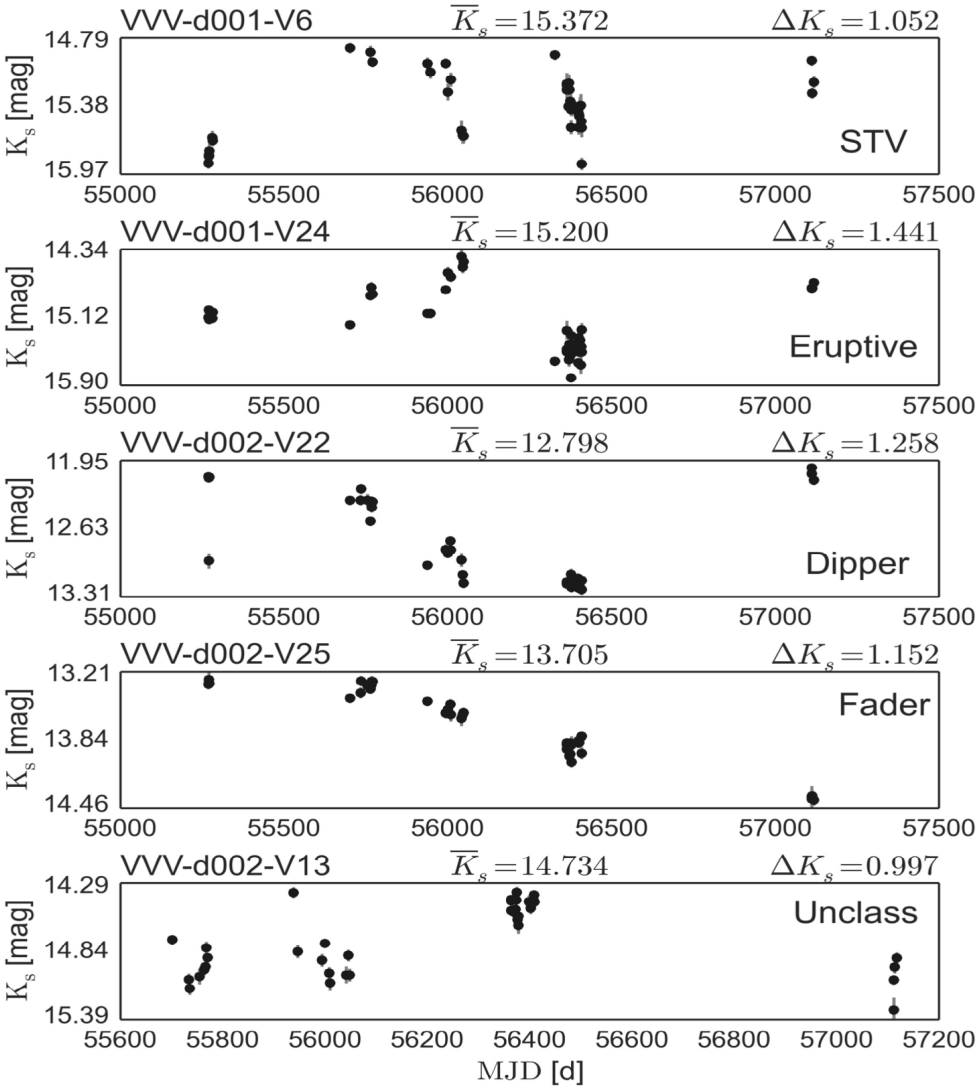


Figure 5.6: Examples of K_s-band time series of irregular variables inside the SFR1. At the top of each plot, the identification ID, mean magnitude K_s, amplitude ΔK_s, and variable type are shown.

29 OB type, Wolf-Rayet, and emission-line stars, adding 18 newly identified objects to the existing catalogs.

Later, in Medina et al. (2021) we reported variables stars found in SFR2. A total of 4,825,964 point sources were detected, of them 3,570,064 sources had more than 20 epochs and an amplitude greater than 0.2 magnitudes, and their time series were constructed and analyzed.

Under the constraints for the selection of Irregular variable stars, we found 196 variable sources. Specifically, in this FoV, we selected sources with $\Delta K_s >$ 0.6 mag and η values lower than 0.95. The relatively small number of selected objects compared to a large number of analyzed time series indicates that our sample's is under very conservative and stringent selection criteria. Nevertheless, we decided to report only highly reliable variable sources rather than a list containing all possible candidates.

The second panel (from top to bottom) of figure 5.3 shows the ΔK_s distribution for the sources found in the FoV as a function of the mean K_s magnitude. In the background, as gray hexagons, the distribution of amplitudes of the whole tile d046 is shown for comparison. The amplitude interval is between $0.661 < \Delta K_s < 3.521$ mag with an average value of $\langle \Delta K_s \rangle = 1.22$ mag, which are similar values than in SFR1.

Regarding the time series profiles, we find 47 irregular variable sources with eruptive characteristics, 25 dippers, 17 faders, 24 LAE, 37 short-term variables, and 45 sources that remain unclassified. Examples of these categories are shown in Figure 5.8 and 5.9. However, it is necessary to mention that some stars may have alternative classifications: d045_v4 and d046_v26 could be faders; d083_v7 could be a dipper; d084_v12 could be an STV, because short-term scatter has the highest amplitude; and d045_v24 and d083_v14 could be eruptive variables.

About the spatial distribution: The location of the identified irregular variable sources inside SFR2 can be seen on the panoramic view of the G305 star-forming complex shown in figure 5.7. The false-color view provided by the Spitzer Space Telescope InfraRed Array Camera (IRAC) at mid-infrared wavelengths of 3.6μm to 8μm is in the background. As can be seen, the selected irregular sources extend throughout the whole area. However, the density dis-

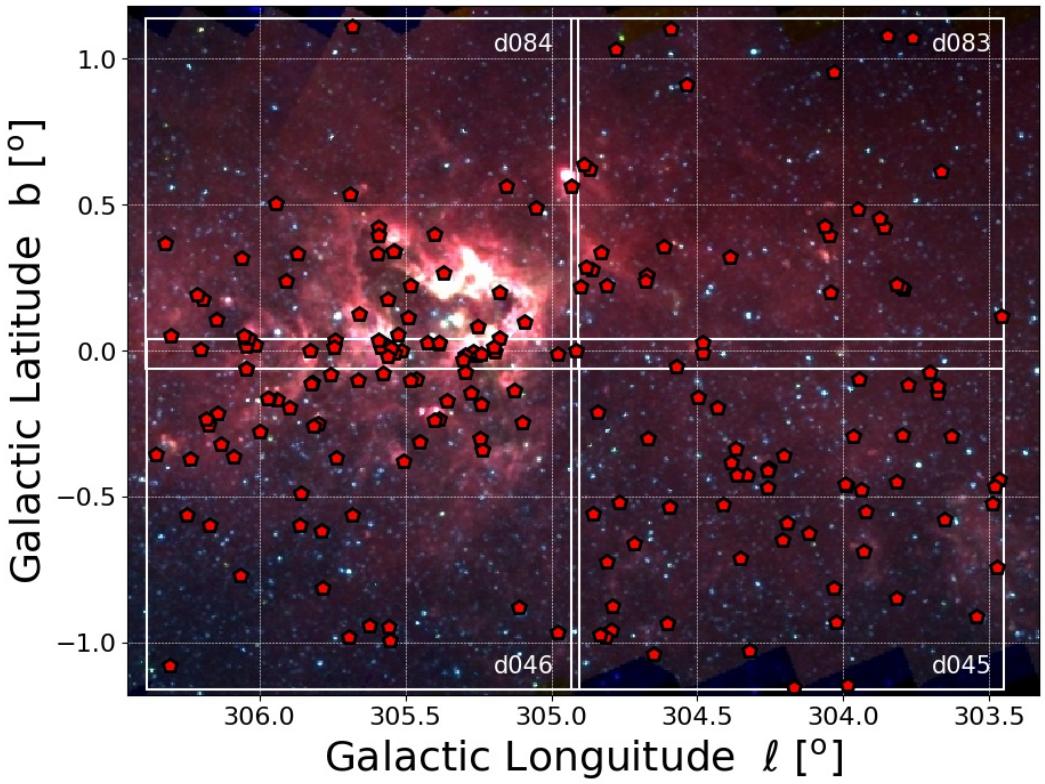


Figure 5.7: The G 305 star-forming complex region. The four observed tiles, named d045, d046, d083, and d084 from the VVV survey are labeled. In the background, the false three-color GLIMPSE Infrared Array Camera (IRAC) (Fazio et al., 2004) image of the complete field is shown, using 8.0μm (red), 4.5μm (green), and 3.6μm (blue). Galactic north is up, galactic east is to the left.

tribution is different: the sources in tiles d046 and d084 are mostly projected on the red filaments across the image. These trace the polycyclic aromatic hydrocarbons (PAH) emission, excited by surrounding interstellar radiation, and become luminescent at wavelengths near 8μm. The PAH emission is typically found in photo-dissociated regions (e.g., Hollenbach & Tielens (1997)).

On the other hand, the sources in the tiles d045 and d083 show a nearly homogeneous distribution. Thus, the distribution of the young sources in G305 traces the ongoing star-formation zones. Some individual examples are shown in Figure 11, where the sources d045_v18 and d045_v35 are projected in the IRAC bands on green nebulosity, indicating an outburst or jet that is bright at 4.5 μm. The right panels show the d084_v6 and d084_v20 sources projected on nebulous regions in 2MASS images.

This result is consisted with (Rice et al., 2015; Contreras Peña et al., 2017b), considering that class I/II objects (compared with class III) are more likely to show variability due to their instability while evolving towards the main-sequence. Thus, one should expect that they are distributed and concentrated in dust and gas over-densities, the trend clearly visible in the G305 region.

Regarding the CCD locus, only a small fraction of the selected sources falls inside the F-region between the reddening vectors projected from the intrinsic color of main-sequence stars and giants. The majority are near the red border of this region and probably are Class III/Class II sources with slight NIR excess. The rest of the sample falls redward of the F-region, thus indicating the presence of intrinsic infrared excess emission, a distinctive feature of young sources with the circumstellar matter.

T-region sources are located redward of the F-region, but blueward of the reddening line projected from the red end of the T Tauri locus. They are considered to be primarily classical T-Tauri stars (Class II objects) with apparent NIR excess. This is consistent with the type of stellar objects we expect to identify, given their IR variable nature. To support this interpretation, we built the membership histogram for all selected sources (second plot and histogram in figure 5.4). About 43% of identified irregular sources are located inside the T-region.

Protostellar sources (i.e., Class I YSOs) are expected to lie redward of the

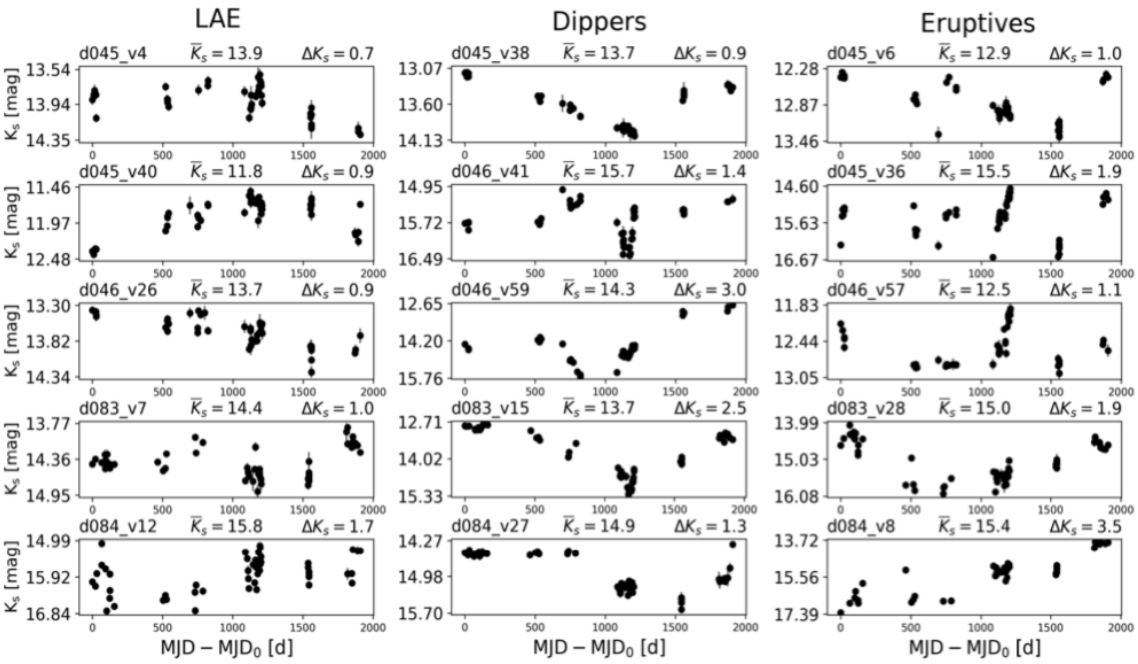


Figure 5.8: Examples of K_s-band time series of irregular variables inside the SFR2. At the top of each plot, the identification ID, mean magnitude K_s, amplitude ΔK_s, and variable type are shown.

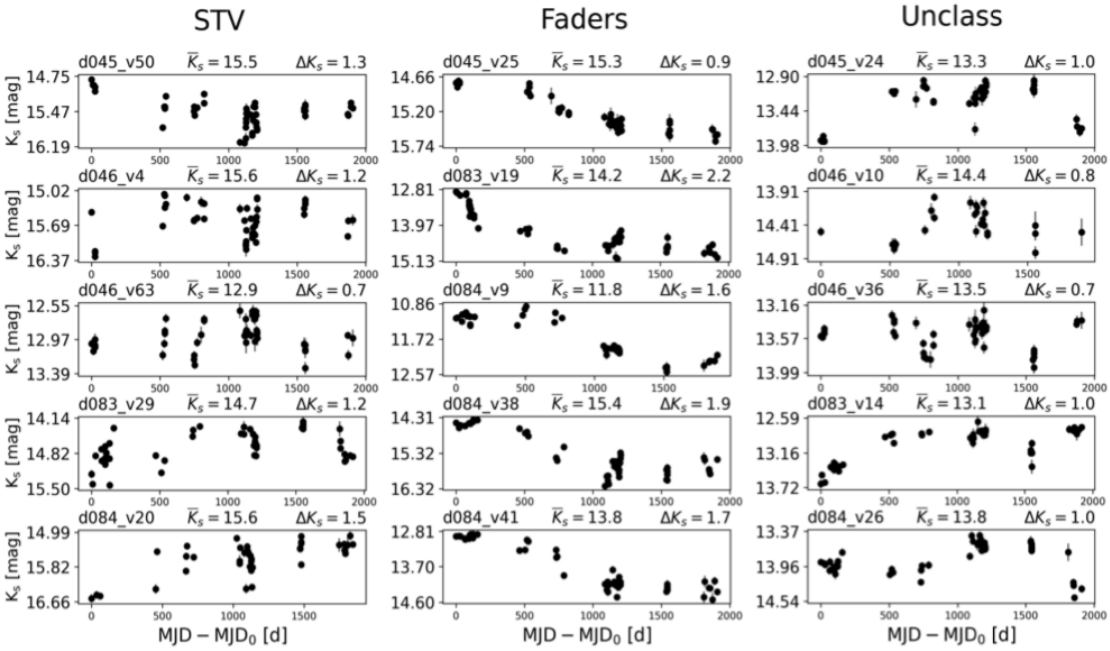


Figure 5.9: More examples of K_s-band time series of irregular variables inside the SFR2.

reddening vector of the classical T Tauri (CTT) locus, in the so-called P-region, given they are intrinsically red due to their evolutionary state. Some 35% of the whole sample falls in this region.

5.4.3 SFR3: NGC 6357 and NGC 6334

The FoV of this star-forming complex is composed of six tiles inside the area of the galaxy's bulge. The number of independent point stars detected N_{SFR3} is 12,943,526 sources, between $10.3 < K_s < 17.4$ mag inside this FoV. Among them, 10,732,063 sources had more than 20 epochs and an amplitude greater than 0.2 magnitudes, and their time series were constructed and analyzed.

The star number N_{SFR3} is around 37.2% greater in comparison to the previous SFR N_{SFR2} (4,825,964 individual sources). This increase is directly related to where in the galaxy we are observing. This fact, in terms of data, translates into a more significant number of outliers or false positives that can be selected by the main selection methods described in the previous section.

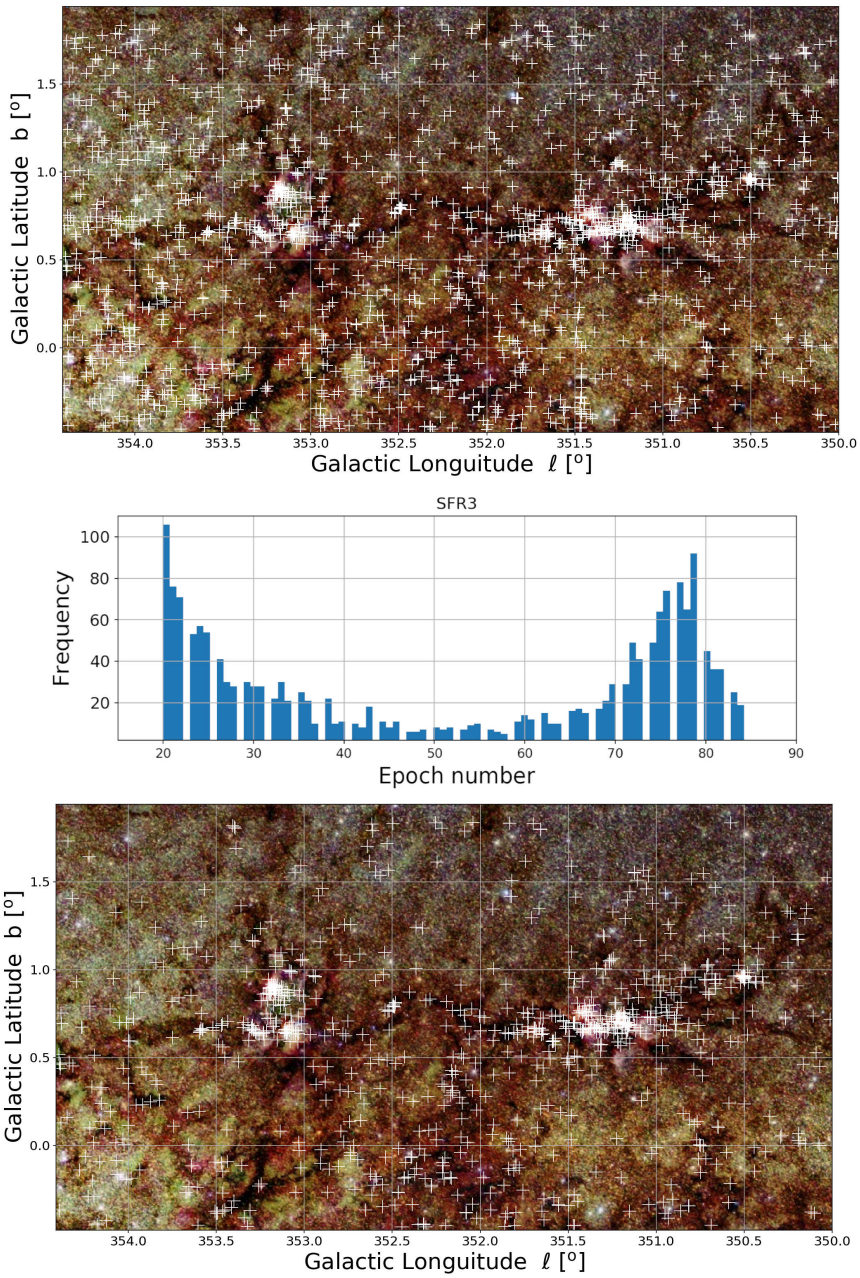


Figure 5.10: Spatial distribution of the Irregular variable sources identified inside the FoV of SFR3. The plot at the top display all the 1825 candidates that fulfill both ΔK_s and η criteria. 2MASS infrared RGB image of the considered field is in the background. Overdensities are located in the filaments and IRDCs at the center of the FoV. Several sources are distributed homogeneously. Middle plot: Distribution of the number of the epoch of the 1825 selected sources. A large fraction of sources with low epoch numbers can be contaminants and only will be considered sources with more than 50 epochs. Bottom plot: Spatial distribution of the 765 sources left after the extra filter.

In the first instance, the time series were analyzed the same way as in SFR1 and SFR2. Therefore, 1825 sources that fulfilled the high amplitude irregular source criteria were chosen. The distribution of their amplitudes can be seen in the third plot of the image 5.3. In this plot, it is possible to see the increase in uncertainty, reflected in the error bars of ΔK_s in several selected sources with large amplitude.

Figure 5.10 displays a false-color image composition using 2MASS survey at the background, which reveals several features associated to massive star formation, some of them also visible in the DSS visual image 5.10. By looking at the spatial distribution of the 1825 sources, we can see that the distribution of the selected sample is concentrated just where we would expect: at the very center of both NGC 6334 (Cat's Paw Nebula, right side) and NGC 6357 (left side) where the massive dark filaments converge, and in the "star-forming axis" structure, where it is possible to find several HII regions formed by massive stars.

However, it is also possible to find selected sources around this massive star-forming complex, where many stars were located in areas of the tiles that look suspicious. The histogram between the FoVs in figure 5.10 displays the distribution of the number of epochs within the selected sample in SFR3, where the inflection point between 45-60 epochs is evident. This is why we decided to make a new filter based on the number of observations. Conservatively, we have chosen the sources with epochs greater than or equal to 50 observations to ensure the selection.

After applying the epoch filter, there are still 765 sources left, displayed in the spatial distribution at the bottom plot of image 5.10. The overdensities around HII regions remain, as well as those located within massive filaments and IRDCs inside the FoV. At the same time, many sources around the filamentary structures have disappeared.

Even so, the previous filter was insufficient to eliminate possible saturated stars, relatively close sources, bogus or other types of outliers. Hence, the individual check of the time series of each star was necessary, and the corrob-

oration of the variable stars in the paw-print images in the VSA page[1] was performed. After this final filter, 725 sources are left in this FoV.

Regarding the overdensities around the center of NGC 6334 ($\ell[^\circ]$,$b[^\circ]$) = (351.140,0.463) and NGC 6357 ($\ell[^\circ]$,$b[^\circ]$) = (353.166,0.890), it is interesting to see that irregular variable sources are concentrated at the edges of the expanding HII regions shaped by the massive stars, especially in the areas among the ionizing radiation fronts, where the layers of interstellar material have been compressed. These might be the conditions to generate induced collapse of a high and low-mass star, and the detected over-densities suggest triggered star formation.

Other minor over densities of more than 4 objects were identified scattered around the massive filament and IRDCs, around the coordinates ($\ell[^\circ]$, $b[^\circ]$)=(350.502, 0.958), ($\ell[^\circ]$, $b[^\circ]$)=(350.705, 1.025), ($\ell[^\circ]$, $b[^\circ]$)=(351.690, 0.667), and around the cluster called G3CC 47, with coordinates ($\ell[^\circ]$, $b[^\circ]$)=(352.488, 0.796).

An unexpected consequence of the many epochs inside the time series of the irregular variable stars is the difficulty of categorizing the time series in terms of the morphology described in 5.2. The time series generally presents all the features described in their different windows of high cadence. When it was possible, the irregular variable star was classified using the scheme, but the only recognizable behavior in time of the variables was the faders/dippers morphology due to their monotonic decreasing/increasing magnitude.

The membership histogram associated to SFR3 displayed on the right part of image 5.11 reveals that aprox. 35 % of the variables are projected en T region.

The upper panels in figure 5.12 display false-color image compositions using the VVV J-band (blue), the H-band (green), and the first epoch (2010) and last epoch (2015) of the K_s-band (red). The circle in the middle of both images points to the variable SFR3_521, whose color is identified in the "P" category, located just where we expect an intrinsically reddened star: inside an IRDC. SFR3_521 is associated with NGC 6357 SFR. Their time series (the bottom panel) shows abrupt changes in its luminosity in short time windows with

[1] `http://horus.roe.ac.uk:8080/vdfs/VgetImage_form.jsp`

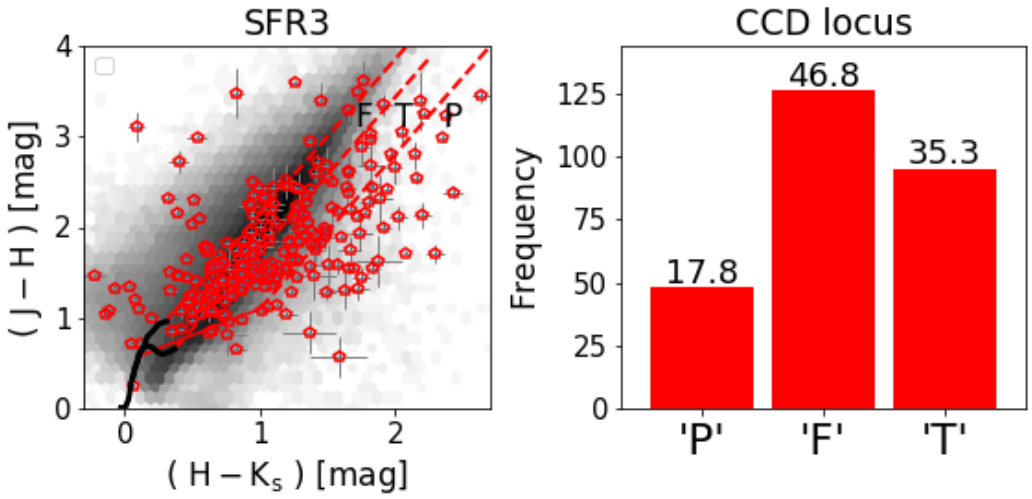


Figure 5.11: Left: The color-color diagram of irregular variable stars identified inside SFR3 FoV. In the background, as gray dots, the distribution of tile b341. Right: The histogram of F, T and P classification inside the CCD.

some scatter along the primary trend of variability, showing the amplitude of $\Delta K_s = 2.893 \pm 0.056$ magnitudes. With this scene, the variable SFR3_521 could be a great candidate for a massive YSO and help us to identify other massive dense cores of high-mass star formation in their vicinity.

5.4.4 SFR4: The Lagoon Nebula Complex

The SFR4 was composed of four tiles located at the very bulge of the galaxy, just about 6.3 degrees away from the very center. The effects of the stellar crowding are measurable given the total number of $N_{SFR4} = 11,236,487$ point sources detected. Magnitudes between $10.8 < K_s < 17.2$ were measured inside this FoV. Only 9,407,898 sources had more than 20 epochs and an amplitude greater than 0.2 magnitudes, and their time series were constructed and analyzed.

Using the main selection methods, 1091 sources were selected.

Obvious overdensities are distributed over the Lagoon nebula, near the massive central star that lights up the emission nebula. In the same way, there

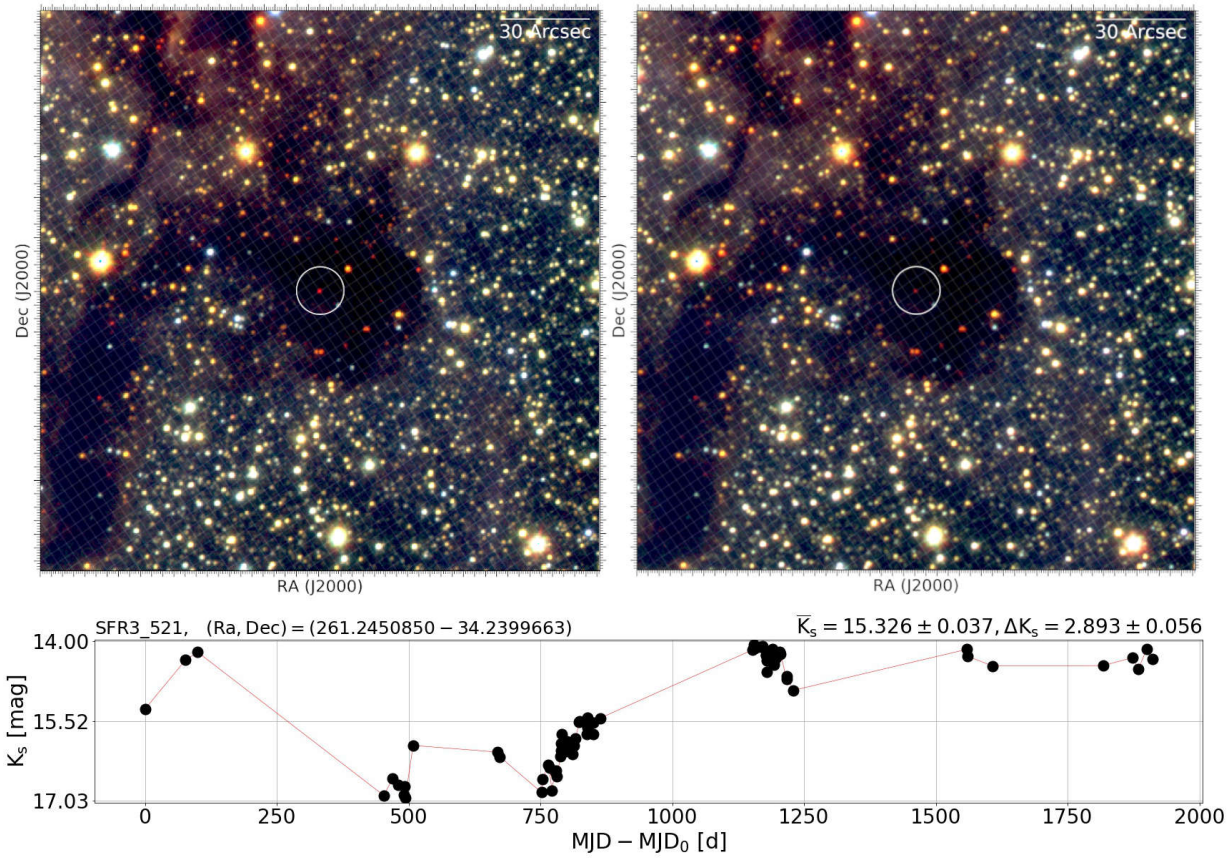


Figure 5.12: The contrasts of the stellar environment of an Irregular Variable star inside the SFR3. Top images: Composed RGB infrared false color image using the VVV JHKs bands. At the center of each image, inside an IRDC, the Irregular Variable star is called SFR3_521 (associated with NGC 6357). Two epochs are considered to visually displays the variability. Bottom: Time series of SFR3_521. Photometric measurements are joined by red straight lines to improve the visualization of the variability over time. The galactic north is up; the galactic east is to the left.

are minor overdensities on the Trifid nebula inside Simeis 188 complex and in dark places between the optical features.

By inspecting the spatial distribution of the selected sample, we have also been able to verify the presence of suspicious sources that could be the product of geometric variations of the VVV tile.

The distribution of observations was considered and found to be similar to SFR3. Therefore only time series with more than 50 epochs were considered. This filter does not affect the variable stars inside b310.

After the individual check, 348 stellar sources remained. The distribution of these stars is in 5.13. Unlike SFR3, the overdensities within the HII regions are more widespread, given the few active O stars associations and locations. Variable stars were located inside the visible range features, and the irregular sources tend to be within the ionized clouds.

The infrared images reveal the distribution of the sources around the IRDCs and the filaments around the FoV. 2MASS image 5.14 of the same FoV indicates the IRDCs distributed in the FoV. Overdensities are located inside these filaments.

This correlation between the IRDCs and the irregular variables is evident in tile b310. The Badee window is characterized by less extinction due to the lack of interstellar dust along the line of sight from Earth, providing a relatively clear view of a portion of the galaxy's bulge.

A large part of the variable stars is located within the dense filament throughout the entire region called Simeis 188. This can be a selection effect due to the number of epochs presented in this tile. But also can be interpreted as a good indicator of the selection of stars.

An overdensity around the star HD 165345 ($\ell[^\circ]$ =6.671, $b[^\circ]$=-1.537) with a reflection nebula around it is observed; 14 sources share the distribution around and have an average amplitude larger than $\Delta K_s > 1.4$ mag, so maybe it is connected with the compression of the material.

Several variable stars are located in the HII region W28 A2, north of the Lagoon nebula. The FoV does not cover all the areas, but it makes sense to have overdensities of this distribution.

The Trifid nebula is crossed by a dense dust filament, mapped by the loca-

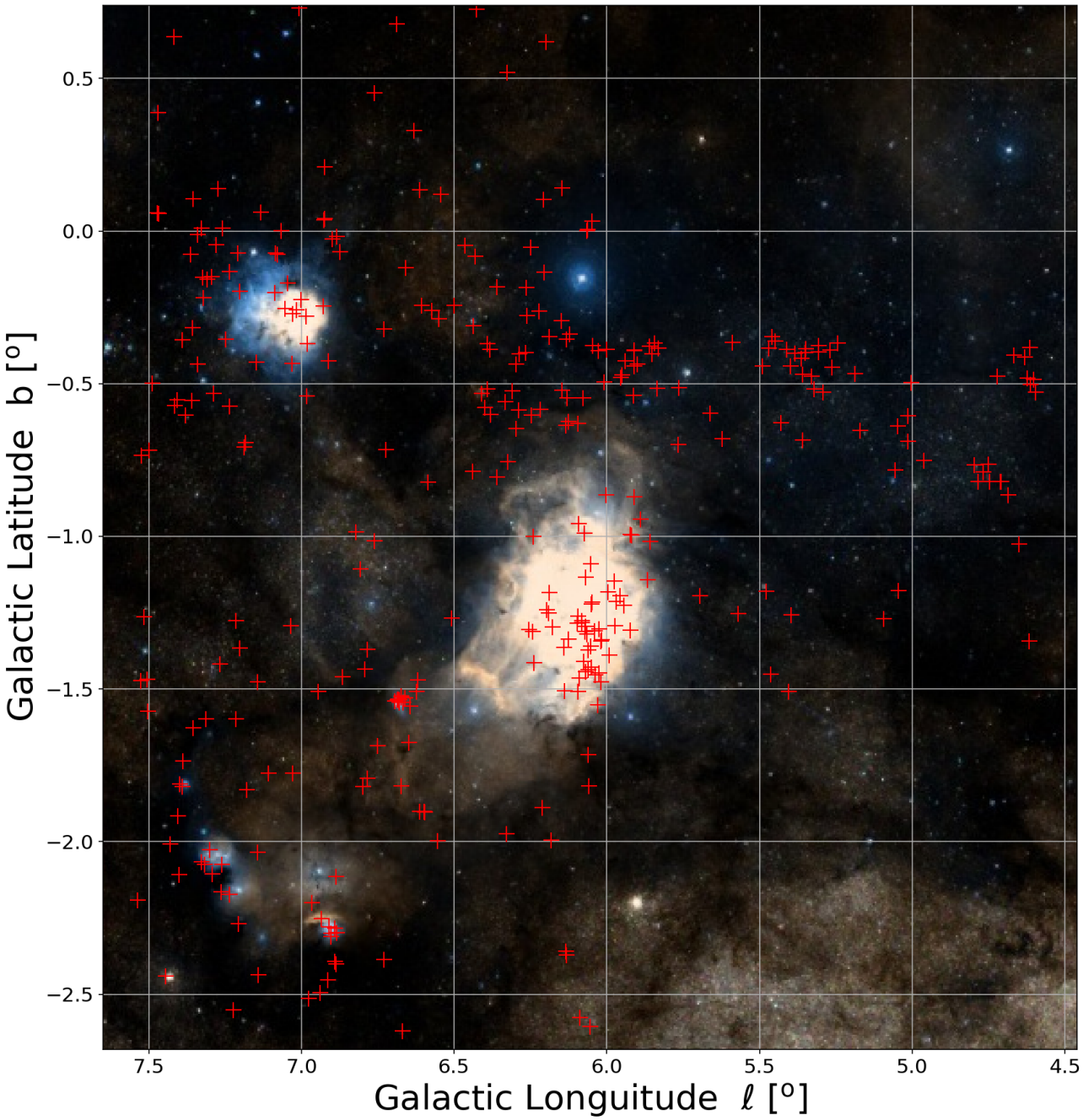


Figure 5.13: Distribution of the selected Irregular Variable objects inside the SFR4. In the background is the DSS optical RGB image of the considered field containing The Lagoon Nebula Complex. The Trifid Nebula (M20) is the northern nebulosity and falls into a small region of the tile b338. The Lagoon Nebula (M8) is covered by tiles b324 and b323. The Simeis 188 region is projected over the high-cadence VVV tile b310. Galactic north is up; galactic east is to the left.

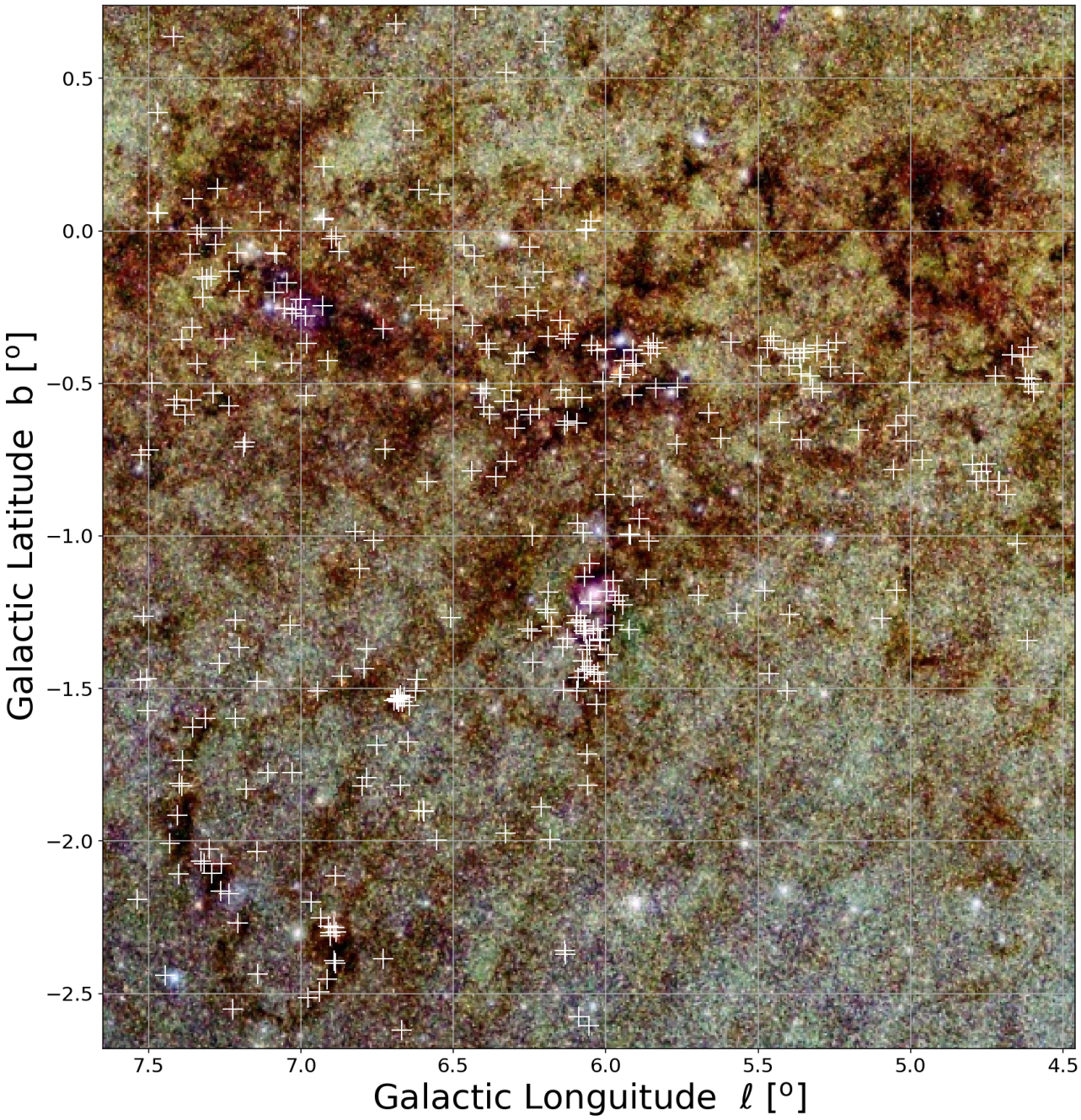


Figure 5.14: Distribution of the selected Irregular Variable objects inside the SFR4. Same FoV than imagen 5.13 but with 2MASS infrared RGB image in background. The visual nebulosities of the Trifid Nebula (M20), the Lagoon Nebula (M8), and Simeis 188 are replaced with dense filaments of cold material, where the selected Irregular Variable objects were located. Galactic north is up; galactic east is to the left.

tion of variable stars. One of these variables, the SFR4_235 (figure 1.5) is associated with an IRDC inside the field, a 1x1 arc sec image of the surroundings of a variable star.

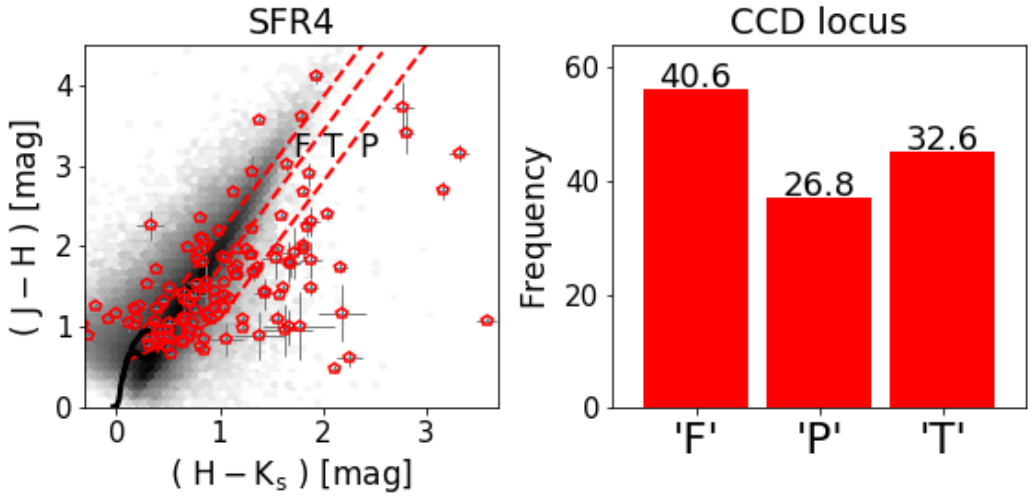


Figure 5.15: Left: The color-color diagram of irregular variable stars identified inside SFR4 FoV. In the background, as gray dots, the distribution of tile b338. Right: The histogram of F, T and P classification inside the CCD.

The membership histogram associated with SFR4 is displayed on the right part of image 5.15 reveals that 40% of variables with measured colors are projected in the "F" region. Intrinsically reddened sources (P and T categories) represent more than 59% of the sample.

5.5 By-product: Periodic stars

An essential and still unexplored population of variable stars are indeed Periodic variable stars. Indeed, their periods is powerful independent tracer of stellar populations properties. For example RR Lyrae traces old (>10 Gyr) stellar populations, Mira stars with different pulsation periods and metallicity can trace SPc of different ages. Short period Miras (P< 300d) usually trace old SPs, while long periods P> 300 are typically younger.

The complete analysis of periodic stars is far from the main objective of this

work. Given the time series from the VVV survey, regular variable sources are detected. During our search for irregular variables, we identified more than 4000 periodic stars with periods between 0.1 days and more than one thousand.

The histogram Fig.5.16 shows the K_s-band amplitude ΔK_s as a function of period P (Bailey's diagram). The amplitudes increase as the period increase, which is consistent with the behavior of pulsating stars. Between 100 and 1000 days are located the so-called "Long Period Variables" (LPVs), which display the lagers uncertainties associated with ΔK_s given the insufficient sampling of the complete periodic cycles registered on our time series.

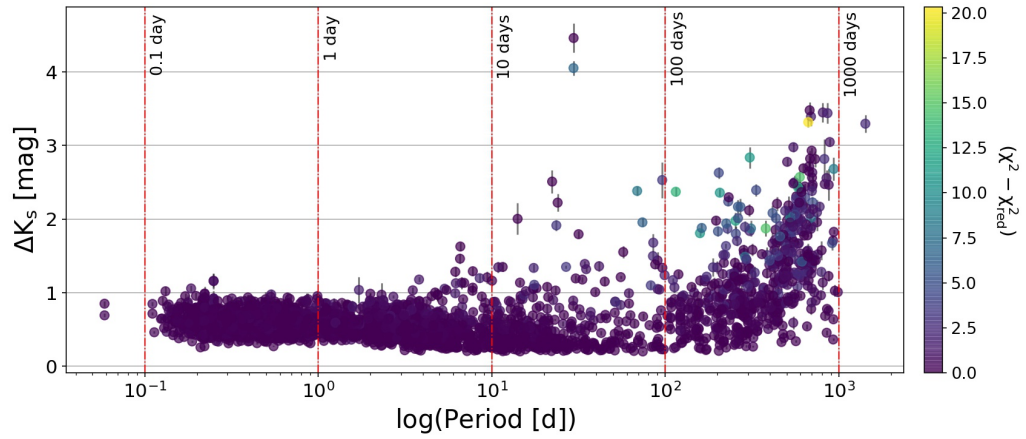


Figure 5.16: K_s-band amplitude ΔK_s as a function of period P (Bailey's diagram) for periodic stars identified in the research FoV.

Long Period Variables are an important part of the periodic sample. In Molina et al. (2018) we reported a new catalog with the variable stars population located near the Galactic center. We detected 353 variable star candidates in a region of $11.5' \times 11.5'$ that comprises two of the most massive young star clusters in our Galaxy: Arches and Quintuplet. Eighty-five percent of these stars (299 objects) have not been reported in previous variability studies carried out in the GC.

The large majority of the variable sources detected are red giant stars, most of them belonging to the AGB phase. Aside from the typical Miras, SR and IRR variables widely described in the literature, we found a sizable sample of

giant stars with long-term trends. We have analyzed the probable causes of these variations.

5.5.1 Variability feature distributions

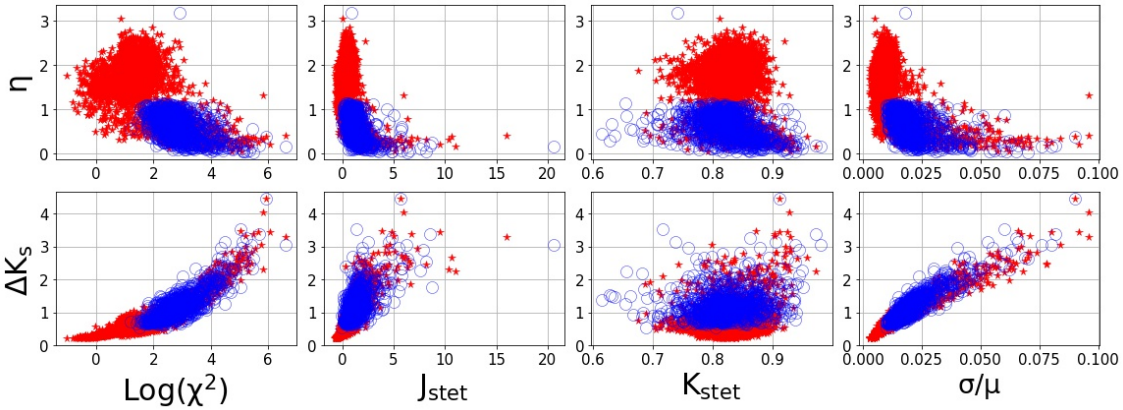


Figure 5.17: Variability feature distributions of periodic (blue open) and irregular (blue) variable stars. The amplitude ΔK_s and η indices are located on the y-axis.

Figure 5.17 shows the Variability feature distributions of the irregular variable stars (red open circles) and periodic variable stars (red stars). The limits and cut-off values are noticeable, as well as the correlation, structures, and extreme values between features.

The number of epochs between the SFRs does not seem relevant in the distribution of selected features, and there are no apparent gaps inside their distribution. In general, the indices identify different structures from the time series and complement each other. Short-period pulsating stars tend to have low amplitudes, and the η index interprets as a star with Gaussian-distributed variability. From this point of view, it makes sense that Irregular Variable sources populate different places compared to periodic variable stars. Several regular stars, such as RBGs and LPVs, display large ΔK_s and share the distribution locations with the irregular variable stars.

Therefore, with enough information and features (with more extensive knowledge), it could be possible to detect relevant and strong correlations

among the features, which can be understood through their distributions. This information is fundamental to researching many aspects of time series, finding multidimensional structures, and increasing the probability of correctly classifying the nature of a specific source.

CHAPTER 6

Summary

We present the systematic research and selection of a sample of 1096 high-amplitude irregular variable stars distributed along the bulge and the south-plane of the galaxy.

Using the time-series photometry in the K_s bandpass provided by the VVV survey for the sources in this region, we generated and analyzed more than 24 million unevenly sampled time series over five years between 2010 and 2015. An automated process was developed to obtain the K_s-band photometric measurement directly from the VVV tiles images. PSF photometry was performed, calibrated to the standard system, and constructed the time series of every independent point source with more than 20 observations and an amplitude larger than 0.2 magnitudes.

We choose four different star-forming regions for this analysis: IC 2944/2948 (SFR1); G305 (SFR2); NGC 6357/NGC 6334 (SFR3), and the Lagoon nebula complex (SFR4) to detect real candidates with prominent NIR variability. We analyzed 1411 VVV tiles images, with more than 1,12 Terabytes of information and more than 25 million point sources in all the considered FoV.

We examined several parameters behind time series to characterize the variability, such as the profile in time, their NIR color, and spatial distribution.

We analyzed and tested several variability indices (features) and concluded that the total amplitude ΔK_s and the η index can be used as principal variability indicators for YSOs variability.

The Irregular Variable sample presents amplitudes $\Delta K_s > 0.6$ mag and η values < 0.95 (except for a variable source inside the SFR1 with $\eta = 3.2$). These thresholds allowed us to select reliable variable star candidates with low content of contaminant sources, such as saturated objects or bogus.

Of the 1096 high-amplitude irregular variable objects, 524 present photometric measurements in J- and H-bands allowing us to construct the color-color diagram and quantify their position inside their CCD. More than 58% of the sources with these measurements present location related to intrinsic reddening; therefore, their variability can be connected to the early stages of the YSOs (stages I-II).

The spatial location of classified variables correlates with star-formation features in both visible and infrared bands. Overdensities of the sample are distributed over HII regions, IRDCs, and their surroundings.

We identified more than 3000 periodic stars inside the VVV time series analyzed as a byproduct.

Bibliography

Alcock C. et al., 1993, in Astronomical Society of the Pacific Conference Series, Vol. 43, Soifer B. T., ed, Sky Surveys. Protostars to Protogalaxies, p. 291

Alonso-García J., Dékány I., Catelan M., Contreras Ramos R., Gran F., Amigo P., Leyton P., Minniti D., 2015, AJ, 149, 99

Alonso-García J., Mateo M., Sen B., Banerjee M., Catelan M., Minniti D., von Braun K., 2012, AJ, 143, 70

Andre P., Ward-Thompson D., Barsony M., 2000, Protostars and Planets IV, 59

Arnaboldi M., Neeser M. J., Parker L. C., Rosati P., Lombardi M., Dietrich J. P., Hummel W., 2007, The Messenger, 127

Arnaboldi M. et al., 2012, The Messenger, 149, 7

Audard M. et al., 2014, Protostars and Planets VI, 387

Baade W., 1946, PASP, 58, 249

Barbá R. H. et al., 2015, A&A, 581, A120

Bessell M. S., Brett J. M., 1988, PASP, 100, 1134

Bolte M., 1987, ApJ, 319, 760

Bonnarel F. et al., 2000, A&AS, 143, 33

Bonnell I. A., Vine S. G., Bate M. R., 2004, MNRAS, 349, 735

Borissova J. et al., 2011, A&A, 532, A131

Borissova J. et al., 2014, A&A, 569, A24

Borissova J. et al., 2016, AJ, 152, 74

Borissova J. et al., 2019, AJ, 158, 46

Caratti o Garatti A. et al., 2017, Nature Physics, 13, 276

Carpenter J. M., Hillenbrand L. A., Skrutskie M. F., 2001, AJ, 121, 3160

Casali M. et al., 2007, A&A, 467, 777

Catelan M. et al., 2011, in McWilliam A., ed, RR Lyrae Stars, Metal-Poor Stars, and the Galaxy, p. 145

Cernicharo J. et al., 1998, Science, 282, 462

Chibueze J. O. et al., 2014, The Astrophysical Journal, 784, 114

Clark J. S., Porter J. M., 2004, A&A, 427, 839

Cody A. M. et al., 2014, AJ, 147, 82

Conti P. S., Crowther P. A., 2004, MNRAS, 355, 899

Contreras Peña C. et al., 2017a, MNRAS, 465

Contreras Peña C. et al., 2017b, MNRAS, 465, 3011

Contreras Peña C. et al., 2017c, MNRAS, 465, 3011

Cross N. J. G. et al., 2012, A&A, 548, A119

Dalton G. B. et al., 2006, in Proc. SPIE, Vol. 6269, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, p. 62690X

Damiani F., Prisinzano L., Micela G., Sciortino S., 2019, A&A, 623, A25

Davies B., Hoare M. G., Lumsden S. L., Hosokawa T., Oudmaijer R. D., Urquhart J. S., Mottram J. C., Stead J., 2011, MNRAS, 416, 972

Dékány I. et al., 2015, ApJ, 812, L29

Drake A. J. et al., 2009, ApJ, 696, 870

Elorrieta F. et al., 2016, A&A, 595, A82

Evans I., Neal J. et al., 2009, ApJS, 181, 321

Eyer L., Mowlavi N., 2008, in Journal of Physics Conference Series, Vol. 118, Journal of Physics Conference Series, p. 012010

Fazio G. G. et al., 2004, ApJS, 154, 10

Fazio G. G. et al., 2004, The Astrophysical Journal Supplement Series, 154, 10

Ferreira Lopes C. E. et al., 2020, MNRAS, 496, 1730

Frost A. J., Oudmaijer R. D., de Wit W. J., Lumsden S. L., 2021, arXiv e-prints, arXiv:2102.05087

Fukui Y. et al., 2018, PASJ, 70, S41

Gran F. et al., 2016, A&A, 591, A145

Green D. A., 2014, Bulletin of the Astronomical Society of India, 42, 47

Guo Z. et al., 2020, MNRAS, 492, 294

Guo Z. et al., 2019, Monthly Notices of the Royal Astronomical Society, 492, 294

Herbig G. H., 1966, Vistas in Astronomy, 8, 109

Herbig G. H., 1989, in European Southern Observatory Conference and Workshop Proceedings, Vol. 33, Reipurth B., ed, European Southern Observatory Conference and Workshop Proceedings, p. 233

Hillenbrand L. A., Strom S. E., Vrba F. J., Keene J., 1992, ApJ, 397, 613

Hindson L., Thompson M. A., Urquhart J. S., Faimali A., Clark J. S., Davies B., 2012, MNRAS, 421, 3418

Hoffman D. I., Harrison T. E., McNamara B. J., 2009, AJ, 138, 466

Hollenbach D. J., Tielens A. G. G. M., 1997, ARA&A, 35, 179

Huijse P., Estevez P. A., Zegers P., Principe J. C., Protopapas P., 2011, IEEE Signal Processing Letters, 18, 371

Jenness T. et al., 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9913, Chiozzi G., Guzman J. C., eds, Software and Cyberinfrastructure for Astronomy IV, p. 99130G

Kaiser N. et al., 2002, in Proc. SPIE, Vol. 4836, Tyson J. A., Wolff S., eds, Survey and Other Telescope Technologies and Discoveries, p. 154

Kalari V. M., 2021, ApJ, 921, 81

Kohoutek L., Mayer P., Lorenz R., 1999, A&AS, 134, 129

Krabbendam V. L., Sweeney D., 2010, in Proc. SPIE, Vol. 7733, Ground-based and Airborne Telescopes III, p. 77330D

Lada C. J., 1987, in IAU Symposium, Vol. 115, Peimbert M., Jugaku J., eds, Star Forming Regions, p. 1

Lada C. J., Gull T. R., Gottlieb C. A., Gottlieb E. W., 1976, ApJ, 203, 159

Law N. M. et al., 2009, PASP, 121, 1395

Li Z. Y., Banerjee R., Pudritz R. E., Jørgensen J. K., Shang H., Krasnopolsky R., Maury A., 2014, in Beuther H., Klessen R. S., Dullemond C. P., Henning T., eds, Protostars and Planets VI, p. 173

Lima E. F., Bica E., Bonatto C., Saito R. K., 2014, A&A, 568, A16

Lomb N. R., 1976, Ap&SS, 39, 447

Majewski S. R. et al., 2017, AJ, 154, 94

McKee C. F., Ostriker E. C., 2007, ARA&A, 45, 565

McSwain M. V., Gies D. R., 2005, ApJS, 161, 118

Medina N. et al., 2018, ApJ, 864, 11

Medina N. et al., 2021, The Astrophysical Journal, 914, 28

Meyer M. R., Calvet N., Hillenbrand L. A., 1997, AJ, 114, 288

Minniti D. et al., 2017, AJ, 153, 179

Minniti D. et al., 2010, , 15, 433

Molina C. N., Borissova J., Catelan M., Lucas P. W., Medina N., Contreras Peña C., Kurtev R., Minniti D., 2018, Monthly Notices of the Royal Astronomical Society, 482, 5567

Mottram J. C. et al., 2011, ApJ, 730, L33

Nishiyama S. et al., 2006, ApJ, 638, 839

Ogura K., Sugitani K., Pickles A., 2002, AJ, 123, 2597

Ojha D. K. et al., 2004, ApJ, 616, 1042

Parzen E., 1962, The Annals of Mathematical Statistics, 33, pp. 1065

Perryman M. A. C., 2005, in Astronomical Society of the Pacific Conference Series, Vol. 338, Seidelmann P. K., Monet A. K. B., eds, Astrometry in the Age of the Next Generation of Large Telescopes, p. 3

Press W. H., Rybicki G. B., 1989, ApJ, 338, 277

Principe J., 2010, Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives, Information Science and Statistics. Springer New York

Prisinzano L. et al., 2019, A&A, 623, A159

Quanz S. P., Henning T., Bouwman J., van Boekel R., Juhász A., Linz H., Pontoppidan K. M., Lahuis F., 2007, ApJ, 668, 359

Rebull L. M. et al., 2014a, AJ, 148, 92

Rebull L. M. et al., 2014b, AJ, 148, 92

Reipurth B., 2008, Handbook of Star Forming Regions, Volume II: The Southern Sky, Vol. 5

Reipurth B., Corporon P., Olberg M., Tenorio-Tagle G., 1997, A&A, 327, 1185

Rho J., Lefloch B., Reach W. T., Cernicharo J., 2008, in Reipurth B., ed, Handbook of Star Forming Regions, Volume II, Vol. 5. p. 509

Rice T. S., Reipurth B., Wolk S. J., Vaz L. P., Cross N. J. G., 2015, AJ, 150, 132

Rieke G. H., Lebofsky M. J., 1985, ApJ, 288, 618

Russeil D., Adami C., Bouret J. C., Hervé A., Parker Q. A., Zavagno A., Motte F., 2017, A&A, 607, A86

Russeil D. et al., 2016, Astronomy Astrophysics, 587

Russeil D., Zavagno A., Motte F., Schneider N., Bontemps S., Walsh A. J., 2010, A&A, 515, A55

Rényi A., 1961, in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. University of California Press, Berkeley, Calif., p. 547

Sadaghiani M. et al., 2020, A&A, 635, A2

Saito R. et al., 2010, The Messenger, 141, 24

Saito R. K. et al., 2012, A&A, 537, A107

Sana H., James G., Gosset E., 2011, Monthly Notices of the Royal Astronomical Society, 416, 817

Scargle J. D., 1982, ApJ, 263, 835

Schechter P. L., Mateo M., Saha A., 1993, PASP, 105, 1342

Schmeja S., Klessen R. S., 2006, A&A, 449, 151

Shannon C. E., 1948, The Bell System Technical Journal, 27, 379

Shimazaki H., Shinomoto S., 2006, in , p. 1289

Shin M.-S., Sekora M., Byun Y.-I., 2009, MNRAS, 400, 1897

Smith L. C. et al., 2018, MNRAS, 474, 1826

Sokolovsky K. V. et al., 2017, MNRAS, 464, 274

Stetson P. B., 1987, PASP, 99, 191

Stetson P. B., 1996, PASP, 108, 851

Sugitani K. et al., 2002, ApJ, 565, L25

Taylor M. B., 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Shopbell P., Britton M., Ebert R., eds, Astronomical Data Analysis Software and Systems XIV, p. 29

Taylor M. B., 2006, in Astronomical Society of the Pacific Conference Series, Vol. 351, Gabriel C., Arviset C., Ponz D., Enrique S., eds, Astronomical Data Analysis Software and Systems XV, p. 666

Teixeira G. D. C. et al., 2018, A&A, 619, A41

Thackeray A. D., Wesselink A. J., 1965, MNRAS, 131, 121

Thompson R., 1985, Journal of the Royal Statistical Society. Series B (Methodological), 47, 53

Tothill N. F. H., Gagné M., Stecklum B., Kenworthy M. A., 2008, in Reipurth B., ed, Handbook of Star Forming Regions, Volume II, Vol. 5. p. 533

Udalski A., 2003, , 53, 291

Vanderplas J., Connolly A., Ivezić Ž., Gray A., 2012, in Conference on Intelligent Data Understanding (CIDU), p. 47

VanderPlas J., Connolly A. J., Ivezic Z., Gray A., 2012, in Proceedings of Conference on Intelligent Data Understanding (CIDU), pp. 47-54, 2012., p. 47

Velghe A. G., 1957, ApJ, 126, 302

von Neumann J., 1941a, Ann. Math. Statist., 12, 367

von Neumann J., 1941b, Ann. Math. Statist., 12, 367

Welch D. L., Stetson P. B., 1993, AJ, 105, 1813

Wolk S. J. et al., 2015, AJ, 150, 145

Wolk S. J. et al., 2015, The Astronomical Journal, 150, 145

Zechmeister M., Kürster M., 2009, A&A, 496, 577