Turbulence and Star Formation Efficiency in Giant Molecular Clouds

Raffaele Rani

A thesis submitted in partial fulfilment of the requirements
of Liverpool John Moores University for the degree of
Doctor of Philosophy

July 2021

Declaration

The work presented in this thesis was carried out at the Astrophysics Research Institute, Liverpool John Moores University. Unless otherwise stated, it is the original work of the author.

While registered as a candidate for the degree of Doctor of Philosophy, for which submission is now made, the author has not been registered as a candidate for any other award. This thesis has not been submitted in whole, or in part, for any other degree.

Raffaele Rani
Astrophysics Research Institute
Liverpool John Moores University
IC2, Liverpool Science Park
146 Brownlow Hill
Liverpool
L3 5RF
UK

November 2021

Abstract

The nature of turbulence in molecular clouds is one of the driving factors that influence star formation efficiency. It is speculated that the high star formation efficiency observed in spiral-arm clouds is linked to the prevalence of compressive (curl-free) turbulent modes, while the shear-driven solenoidal (divergence-free) modes appear to be the main cause of the low star formation efficiency that characterises clouds in the Central Molecular Zone (CMZ). Similarly, the analysis of the Orion B molecular cloud confirmed that the dominant solenoidal turbulence is compatible with its low star formation rate.

However, turbulent modes vary locally and at different scales within the cloud, and turbulent motions surrounding the main star-forming regions display a strongly compressive nature. This evidence points to inter-and intra-cloud fluctuations of the solenoidal modes being an agent for the variability of star formation efficiency and cloud collision being a facilitator of stars' formation through the production of highly compressive gas flows.

This thesis presents a quantitative estimation of the relative fractions of momentum density in the solenoidal and compressible modes of turbulence in the plane molecular clouds found in the ^{13}CO/C^{18}O ($J = 3 \rightarrow 2$) Heterodyne Inner Milky Way Plane Survey (CHIMPS). This calculation is achieved through a statistical method that allows us to reconstruct the 3-dimensional distribution of the density momentum from its line-of-sight projected counterparts (zeroth, first, and second velocity moments) provided by the observations, producing an estimate of the power contained in the solenoidal and compressive turbulent modes within each cloud.

The project investigates how different fractions of compressive and solenoidal modes in CHIMPS clouds probe the variation of the star formation efficiency across clouds

with varying environments. A negative correlation between the solenoidal fraction and star formation efficiency is found. This feature is consistent with the hypothesis that solenoidal modes prevent or slow down the collapse of dense cores. In addition, the relative power in the solenoidal modes of turbulence (solenoidal fraction) appears to be higher in the inner Galaxy declining with a shallow gradient with increasing Galactocentric distance. Outside the Inner Galaxy, the slowly, monotonically declining values of the solenoidal fraction suggest that the solenoidal fraction is unaffected by the spiral arms.

The sample of clouds considered is extracted via the dendrogram-based Spectral Analysis for Interstellar Molecular Emission Segmentation (SCIMES).

The comparison of the geometrical and physical properties of the SCIMES extracted ^{13}CO (3-2) clouds in CHIMPS with the results originally obtained with the FellWalker method show that the SCIMES segmentation includes a wider range of cloud sizes. In crowded fields, SCIMES produces more detailed maps of the structure of molecular cloud, by identifying and tracing out more rarefied features and avoiding "clump localisation" with artificial boundaries arising in the FellWalker extraction. The physical properties defined by the volume and mass of individual clouds mirror this feature. The survey-wide distributions of physical properties of the ^{13}CO emission however are similar in the two segmentations. To compare the properties of the extracted clouds to those identified using a different tracer, a SCIMES segmentation of the $^{12}\mathrm{CO}(3-2)$ emission from the CO High Resolution Survey (COHRS) through SCIMES is considered (where the data are available).

Acknowledgements

First of all, I would like to thank my supervisor Toby Moore for all of his efforts and his patience. I am profoundly grateful to David Eden for his supervision and his availability. It has been a privilege to be part of your team at ARI. You brought me from zero to this thesis in two years! I look forward to new projects together. I thank Ivan Baldry for being on the project and completing the work on compact galaxies, which started with my and Trisha's theses, while I was focusing on molecular clouds. I am very grateful to Steven Longmore at ARI, Andrew Rigby, Sarah Ragan and Matt Smith at Cardiff University and the SCIMES team (Ana Duarte, Eric Rosolovski and Dario Colombo). A special thank you goes out to Harriet Person, Kevin Silva, and Miriam Fuchs at EAO who welcomed me to Hilo and guided me during my observation run at JCMT. I thank Phil James for being my internal examiner and for coordinating the first LJMU distance learning course which formally started my adventure in Astrophysics. I thank the Doctoral Academy for financing my PhD studies through an LJMU scholarship. Thank you, Andreea Font and Ian McCarthy for supporting my application. Alexia Montaubin at FET, Anna Hodgkinson and Maureen Patullo at ARI allow yourself a huge thank you for answering all my enquiries about PhD studies, budget and travel expenses.

Last, I thank my family, friends, gurus, cats and everyone else who is walking this path with me.

Raffaele Rani November 2021

Declaration of Authorship

I, Raffaele Rani, declare that this thesis titled, 'Turbulence and Star Formation Efficiency in Giant Molecular Clouds' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"The night hides the world, but reveals the universe."

- Persian proverb

"Time will never betray dreams, nor will dreams ever betray time."

- Kei and Rei in Captain Harlock: Dimensional Voyage, Leiji Matsumoto

Contents

List of Figures

List of Tables

Abbreviations

ALMA	The **A**tacama **L**arge **M**illimeter **A**rray
APEX	The **A**tacama **P**athfinder **E**xperiment
ATLASGAL	The **A**pex **T**elescope **L**arge **A**rea **S**urvey of the **G**alaxy
CAA	**C**lump **A**ssignment **A**rray
CANFAR	**C**anadian **A**dvanced **N**etwork for **A**stronomical **R**esearch
CfA	**C**enter **f**or **A**strophysics
CHIMPS	The **C**O **H**eterodyne **I**nner **M**ilky **W**ay **P**lan **S**urvey
CMZ	**C**entral **M**olecular **Z**one
COHRS	The **C**O **H**igh **R**esolution **S**urvey
CSC	**C**ompact **S**ource **C**atalogue
CUTEX	The **CU**urvature **T**hresholding **EX**tractor
DGMF	**D**ense **G**as **M**ass **F**raction
FITS	**F**lexible **I**mage **T**ransport **S**ystem
FWHM	**F**ull-**W**idth **H**alf-**M**aximum
GRS	The **G**alactic **R**ing **S**urvey
HARP	The **H**eterodyne **A**rray **R**eceiver **P**rogramme
HMSF	**H**igh **M**ass **S**tar **F**ormation
IMF	**I**nitial **M**ass **F**unction
IR	**I**nfra**R**ed
IRDC	**I**nfra **R**ed **D**ark **C**loud
JCMT	The **J**ames **C**lerk **M**axwell **T**elescope
JPS	The **J**CMT **P**lane **S**urvey
KDA	**K**inematic **D**istance **A**mbiguity
LABOCA	the **L**arge **A**PEX **BO**olometer **CA**amera
LIRG	**L**uminous **I**nfra **R**ed **G**alaxy

LSR	**L**ocal **S**tandard of **R**est
LTE	**L**ocal **T**hermodynamic **E**quilibrium
MALT90	The **M**illimetre **A**stronomy **L**egacy **T**eam 90 GHz Survey
MGPS	**M**opra CO **G**alactic **P**lane **S**urvey
MIR	**M**id **IR**ed
PACS	**P**hotodetector **A**rray **C**amera and textbfSpectrometer
PAWS	**P**dBI **A**rcsecond **W**hirpool **S**urvey
PdBI	**P**lateau **d**e **B**ure **I**nterferometer
PPV	**P**osition-**P**osition-**V**elocity
PDF	**P**robability **D**istribution **F**unction
RMS	The **R**ed **M**sx **S**ource survey
rms	**r**oot-**m**ean **s**quare
SED	**S**pectral **E**nergy **D**istribution
SEDIGISM	**S**tructure **E**xcitation and **D**ynamics of the **I**nner **G**alactic **I**nterstellar **M**edium
SExtractor	**S**ource **E**xtractor
SFE	**S**tar **F**ormation **E**fficiency
SFR	**S**tar **F**ormation **R**ate
SNe	**S**uper **Nova**e
SNR	**S**ignal-to-**N**oise **R**atio
SPIRE	**S**pectral and **P**hotometric **I**maging **RE**icever
ThrUMMS	the **Thr**ee-mm **U**ltimate **M**opra **M**ilky Way **S**urvey
ULIRG	**U**ltra **L**uminous **I**nfra **R**ed **G**alaxy
UV	**U**ltra**V**iolet
YSO	**Y**oung **S**tellar **O**bject
ZAMS	**Z**ero **A**ge **M**ain **S**equence

Chapter 1

Introduction

The conversion of molecular gas into stars is one of the fundamental baryonic processes that shape the visible Universe, driving cosmic evolution from the epoch of re-ionisation to present-day Galactic systems.

Despite progress on the broad scenarios of formation and collapse of molecular clouds, the physical processes that initiate and drive the formation of stars and how efficiently they convert gas into stars are still poorly understood. Star formation efficiency (SFE) along with the initial stellar mass function (IMF) are the essential ingredients to construct a predictive model of star formation. To this day, only observations within the Milky Way are able to detect and resolve both gas and stars on size scales of individual star-forming regions. Galactic surveys and single object observations are thus the only means to estimate the relative importance of the physical processes that may impact SFE from local (parsec) scales within individual giant molecular clouds (temperature, turbulence, etc.) to Galaxy-wide scales ($>$ 1 kpc, spiral density wave).

A potential driving agent of star formation has been identified as the relative fraction of turbulence modes in the interstellar molecular gas. In this framework, the high star formation efficiency (SFE) observed in spiral-arm clouds is linked to the prevalence of compressive (curl-free) turbulent modes. In contrast, the low SFE that characterises clouds in the Central Molecular Zone (CMZ) is related to the shear-driven solenoidal (divergence-free) component. The application of statistical methods to the study of turbulence in line-of-sight projected data requires the accurate identification of molecular clouds and their structure in emission maps. A wide range of algorithms has been

devised to 'extract' molecular and their physically significant substructures embedded in the emission. Each of these approaches has its own particular features and performs best under particular circumstances.

These methods are complex and comparing their relative efficiency is often problematic, both because few have been applied to the same dataset and because no common standard of calibration exists. Furthermore, it is exceedingly difficult to cross-correlate the properties of individual clouds between the various catalogues, as these are likely to be defined in different ways.

The work presented in this thesis covers two projects. One aimed to compare two emission extraction algorithms applied to the same CO survey: the dendrogram based Spectral Clustering for Interstellar Molecular Emission Segmentation (SCIMES) and the more well-established watershed FellWalker algorithm. The second project is an attempt to present the first full sample study of the turbulent modes and their relation to SFE in Galactic clouds, thus testing the hypothesis that the SFE depends on the ratio of solenoidal to compressive turbulence within clouds. This has already been suggested for one CMZ cloud and is thought to be consistent with the assumption that the majority of power in SFE variations is concentrated on cloud scales.

1.1 Molecular clouds

The earliest stages of star formation see neutral gas in the the interstellar medium (ISM) aggregating in dense molecular clouds through large-scale hydrodynamic, thermodynamic, or gravitational instabilities.

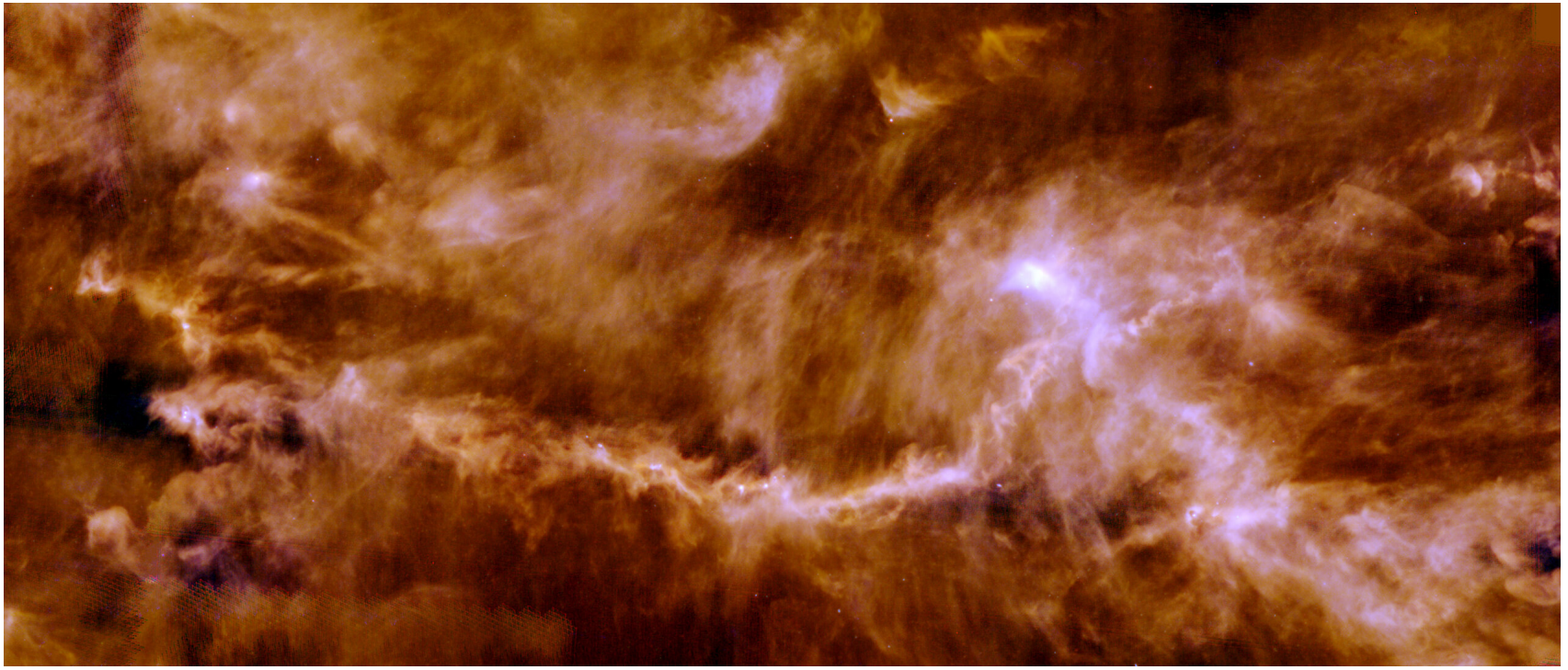


FIGURE 1.1: Gas and dust in the Taurus Molecular Cloud, a giant stellar nursery in the constellation Taurus. The image is a combination of emission detection in three-colour Herschel bands at 160 νm (blue), 250 νm (green) and 500 νm (red), and spans about 5° in length. The data were acquired as part of the Herchel Gould Belt survey Key Programme (André et al., 2010) and published on ESA's Science and Technology website.

These perturbations are associated with colliding, or shearing flows or shocks caused by the gas entering the spiral arms. Dissipative shocks in the supersonic turbulence resulting from the cloud-formation process, then (or concurrently, Heitch et al., 2008) form fragmented, compressed layers, and filaments. Dense fragments become gravitationally self-bound and collapse into the clumps and cores that eventually create stars, while more rarefied structures are transient and dissipate. Since the paradigm characterizing the interstellar medium (ISM) has shifted towards the picture of an inherently dynamic environment, the view of molecular clouds as naturally, largely transient features has succeeded their older characterisation as extant structures in a state of quasi-equilibrium preceding collapse and the formation of stars. One defining characteristic of molecular clouds is that they are not independent isolated entities in space but instead, they are dense condensations in the more widely distributed, mostly atomic gas. Although many molecular clouds appear to have sharp boundaries, these confines do not mark the end of the gas distribution, but they constitute transitions from the molecular gas to the surrounding atomic gas, which forms envelopes of comparable mass (Blitz, 1988). The amount of molecular gas in clouds is predicted to depend on both the local gas density n and the column density N of the material shielding molecular gas from dissociating ultraviolet radiation. These densities also influence the composition of the cloud. A cloud is expected to be predominantly molecular when $N = n^{2/3}$ exceeds a critical threshold that depends on the ultraviolet radiation flux and the dust abundance (Elmegreen, 1989, 1993). Thus, the condition for a cloud to be constituted predominantly by molecular gas does not require that it be gravitationally bound [1]. The molecular content of a cloud (or region of a cloud) could rapidly change because of its sensitivity to varying local conditions (radiation and magnetic fields, for instance, Elmegreen, 1993). Molecular clouds are found in various forms and sizes. The range from small globules (Bok globules) with mass ~ 10 $M_\odot$ contained within ~ 0.5 pc (Clemens & Barvainis, 1988; Clemens et al., 1991), to giant molecular clouds (GMCs) which comprising a total mass of $\sim 10^6$ $M_\odot$ within ~ 100 pc (Roman-Duval et al., 2010) (see also table 1.1). Molecular clouds have highly irregular and complex shapes. Many of them possess wispy filamentary structures that resemble those of atmospheric clouds (see Figure 1.1). The irregular boundaries of molecular clouds found on contour maps show fractal properties (Dickman et al., 1990; Falgarone et al., 1991, 1992; Zimmermann & Stutzki, 1992; Elia et al., 2018). The fractal

[1] The converse holds too: the gas content in a gravitationally bound cloud does not necessarily have to be molecular.

dimension estimated for these clouds presents similar values to those found at various interfaces in turbulent flows (Falgarone et al., 1991; Sreenivasan, 1991; Lee et al., 2016), suggesting that turbulence plays a fundamental role in the formation and evolution of molecular clouds. Commonly, velocity dispersions within molecular clouds are about ten times larger than expected by solely considering thermal properties (Larson, 1981; Rathborne et al., 2009). This is generally interpreted as evidence of turbulence being a prominent factor in creating and sustaining a cloud's internal structure. The complex hierarchical structure (see section 1.2) characterizing molecular clouds is thought to arise as a consequence of complicated interactions of gravity, magnetic fields (Elmegreen & Scalo, 2004; Mac Low & Klessen, 2004; McKee & Ostriker, 2007; L. et al., 2011; Heyer & Brunt, 2012) and supersonic turbulent motions driven at different scales from stellar feedback to Galactic shear (Scalo & Elmegreen, 2004).

Star formation occurs in the densest regions of molecular clouds. The characteristic physical conditions under which these regions collapse are determined by the competition between self-gravity and thermal pressure. This assumption allows us to define the characteristic length (Jeans length, determined by the speed of sound in the gas c_s, the density ρ or number density n of the gas and the gravitational constant G)

$$\lambda_J = \sqrt{\frac{\pi c_s^2}{G\rho}} \sim 2.2\mathrm{pc}\left(\frac{c_s}{0.2\mathrm{kms}^{-1}}\right)\sqrt{\frac{10^2\mathrm{cm}^{-3}}{n}} \tag{1.1}$$

and the corresponding spherical Jeans mass, above which an isothermal fluid parcel collapses under its self-gravity (Draine, 2011)

$$M_J = \frac{4\pi}{3}\rho\left(\frac{\lambda_J}{2}\right)^3 \sim 34M_\odot\left(\frac{c_s}{0.2\mathrm{kms}^{-1}}\right)^3\sqrt{\frac{10^2\mathrm{cm}^{-3}}{n}}. \tag{1.2}$$

played by other physical mechanisms, collapse occurs on timescales ranging from 0.1 Myr for densest cores of $\sim 10^5$ cm^{-3} up to 3 Myr for the regions within GMCs that have volume-averaged densities of $\sim 10^2$ cm^{-3} (Cheavance, 2020). This 'free-fall timescale' (see 4.4.3) is thus a lower bound of the actual collapse time. The early stages of collapse are slowed down by the thermal pressure gradient, magnetic fields (Inoue & Inutsuka, 2012; Vàzquez-Semadeni et al., 2011; Girichidis et al., 2018), turbulence (Klessen et al., 2000; Dobbs & Baba, 2014), Galactic differential rotation through shear and Coriolis

forces (Dobbs & Baba, 2014; Meidt et al., 2018, 2020), and the non-spherical (planar or filamentary) shape of the clouds (Toalà et al., 2012; Pon et al., 2012).

Clearly, star formation is highly disruptive to molecular clouds and the outset of collapse marks a fundamental transition in the evolution of a molecular cloud, leading to it being destroyed or largely restructured (Krumholz, 2019). This transition phase, during which mass gain and mass loss are in approximate equilibrium, may last millions of years before the star-forming cores acquire high enough masses and densities to trigger the formation of (massive) stars (Vàzquez-Semadeni et al., 2017; Krause, 2020). At the end of this stage, the input of energy and momentum from stellar feedback from the newly created star-forming regions (massive stars) becomes prominent and the host cloud is dispersed (Krumholz, 2019; Lopez et al., 2014; Rahner et al., 2017, 2019; Grudič et al., 2018; Haid et al., 2018; Kim et al., 2018; Kruijssen et al., 2019a; McLeod et al., 2020). Stellar feedback consists mainly of three processes: ultraviolet radiation, stellar winds (in the early stage of star formation), and supernovae (SNe). Each mechanism provides a source of energy and momentum that opposes gravity (Krumholz, 2019; Krumholz et al., 2019). The transition between molecular clouds and young stellar regions is rapid, driven by photo-ionisation and stellar winds, it disperses the clouds within a few million years. This cycle is however not universal, but the physical mechanisms controlling the different phases of this process are likely to depend on the environmental conditions. Since the timescales for the formation, internal evolution, and destruction of molecular clouds are all of the same order, these processes cannot be clearly separated in time, and they may all go on simultaneously in different parts of a star-forming complex (Cheavance, 2020).

1.2 Structure of molecular clouds

The distribution and properties of gas within molecular clouds regulate, in part, the characteristics of newly formed stars, their numbers and masses, and the location of star-forming sites. The connection between the features of molecular gas and both the initial mass function and formation rate of new stellar populations has prompted a wide range of theoretical and observational studies geared towards the characterisation of the structure of molecular clouds. Multi-tracer surveys have revealed the hierarchical nature of these structures, showing how high-density, small-scale features are always nested within more rarefied, larger envelopes (Blitz & Stark, 1986; Lada, 1992). This

	Mass ($M_\odot$)	Size (pc)	Density (cm^{-3})	Temperature (K)	Velocity width ($km\,s^{-1}$)
Cloud	$10^2 - 10^6$	$1 - 50$	$50 - 800$	$4 - 8$	$0.3 - 3.0$
Clump	$30 - 10^2$	$0.3 - 3$	$10^2 - 10^4$	$10 - 20$	$0.3 - 3.0$
Core	$0.2 - 30$	0.03-0.2	$10^4 - 10^6$	$8 - 13$	$0.1 - 0.3$

TABLE 1.1: Physical properties of molecular clouds, clumps and cores (Roman-Duval et al., 2009; Bergin & Tafalla, 2007; Dunham et al., 2011; Polychroni et al., 2013). In general, clumps are thought to be the precursors of star clusters, while cores are expected to give rise to individual or multiple star systems.

structural hierarchy is, however, a non-trivial one: at any scale, there appear to be more high density and compact 'clumps' than larger and less dense structures. The densest clumps in a cloud's hierarchy are compact cores, the seeds of star formation. In these regions, over scales of about 0.1 pc (see also table 1.1, the turbulence in the cloud becomes dominated by thermal motions (Goodman et al., 1998; Tafalla et al., 2004; Lada et al., 2008). The physical conditions inside these cores determine the mechanisms that occur in the conversion of molecular gas into stars (di Francesco et al., 2007; Ward-Thompson et al., 2007; Bigiel et al., 2008; Schruba et al., 2011; Urquhart et al., 2018).

At the bottom of the density hierarchy, lie the low-density envelopes that surround the denser regions. The chemical change that characterises the formation of molecular clouds has led to the cataloguing of molecular emission by dividing the interstellar gas into independent, discrete entities. Although this separation provides a useful theoretical distinction between giant molecular clouds and the diffuse multi-phase interstellar medium, it is still unclear whether the density hierarchy continues past this "chemical boundary "(Blitz et al., 2007) extending into the diffuse ISM (Ballesteros-Paredes et al., 1999; Hartmann et al., 2001). In this picture, the molecular phase of the ISM would not be enough to define the bottom of the density hierarchy needed to treat a molecular cloud as an independent, separate entity. This argument is supported by discrepancies between estimated crossing times and expected lifetimes of molecular clouds in some sets of observations. However, the apparent contradictions in the estimated cloud lifetimes in diverse datasets can be reconciled when models of rapid star formation bursts in long-lived clouds (≈ 30 Myr) are considered (Elmegreen, 2007).

Although the hierarchical structure of the ISM continues to large scales past the molecular phase, linking the density of atomic gas in the ISM to molecular clouds is often

difficult, and the analysis of structure within star-forming clouds is restricted to molecular emission. In particular, molecular line emission studies that are to be compared to the 21 cm atomic gas emission are usually affected by degraded spatial resolution due to the long wavelength of the emission. Moreover, fore- and background confusion often makes these studies unreliable. The atomic gas related to molecular clouds can thus only be identified in particular circumstances where either the cloud geometry is known (Pound & Goodman, 1997), self-absorption can be quantified (Li & Goldsmith, 2003a) or a model for photo-dissociation exists (Bensch, 2006). The large spatial dynamic range required in the investigation of the internal structure of molecular clouds constrains useful observation to Galactic samples.

Studies of the Herschel infrared Galactic Plane Survey (Hi-GAL Molinari et al., 2010a) and the Herschel Gould Belt Survey (André et al., 2010), revealed that the morphology of the interiors of molecular clouds is pervaded by networks of filamentary structures (André et al., 2010; Molinari et al., 2010b; Men'shchikov et al., 2010; Arzoumanian et al., 2011). In addition, it was found that the vast majority of star-forming cores reside within filaments (Polychroni et al., 2013; Könyver et al., 2015). The ubiquity of such features, observed in highly-sensitive high-angular resolution submillimetre dust continuum surveys, has rekindled the interest in both observational and theoretical studies on gas flows in filaments.

1.3 Molecular emission

The main species that constitute a molecular cloud are molecular hydrogen, H_2 and inert atomic helium. At the typical temperature of cold ISM ($\sim$ 10 K) emission from these species is practically absent[2]. The next most common molecule in the ISM is carbon monoxide (CO). CO possesses low rotational energy level and radiates at $\sim$ 5 K. Carbon monoxide emission possesses several transitions detectable at millimetre and submillimetre wavelengths. As CO is always associated with the presence of H_2, these features make it an optimal tracer for the observation of molecular clouds (Draine, 2011). The relative CO-to-H_2 abundance can be calculated from the the column density

[2]The H_2 molecule cannot radiate through rotational transitions of the dipole moments as it lacks a permanent dipole moment. Quadrupole transitions also have small transition probabilities and require exceedingly high excitation temperatures ($>$ 500 K) for this molecule to radiate in the cold phase of the ISM (Draine, 2011).

of H_2 (derived from dust extinction or emission, assuming a dust-to-gas ratio, see below) divided by the column density of CO. This ratio, however, is not constant: it depends on the balance between the formation and destruction processes that govern the amount of CO and H_2. Variations in the CO-to-H_2 abundance ratio have been reported in different Galactic environments, in particular the Galactic centre (Sodroski et al., 1995) and outer Galaxy (Brand & Wouterloot, 1995), and in molecular clouds at high Galactic latitudes (Paradis et al., 2012).

In average ISM conditions, CO molecules are most likely to be excited by a both collisions (commonly with H_2) and the absorption of photons. Emission occurs through the quantisation of rotational energy

$$E_J = \frac{\hbar^2}{2I} J(J+1), \tag{1.3}$$

where E_J is the rotational energy of J-th level, $\hbar$ the Planck constant, I the moment of inertia of the molecule[3].

There is a critical density the marks the point at which a molecule's spontaneous emission equals its collision rate with other molecules. This critical density is directly proportional to the collisional cross-section and inversely proportional to the time-averaged velocity of the molecule. Table 1.2 reports values of the critical densities for the most frequent transitions for the most common CO isotopologues (along with emission values and frequencies for the transitions considered).

When CO is denser than this threshold, its energy levels are thermalised, and the gas temperature and column density determine its line intensity. When CO density is below the critical value, its emission intensity is also dependent on the gas volume density. Sub-thermal emission from CO can still occur below the critical density, but it is likely to have very little strength (relative to the column density) (Draine, 2011).

Several CO isotopologues are frequently targeted in millimetre and submillimetre surveys. Since the most abundant ^{12}CO may easily become optically thick, the relatively rarer ^{13}CO and C^{18}O are often observed in millimetre and submillimetre surveys to trace H_2 at higher optical depths. The abundances of these isotopologues with respect to ^{12}CO

[3]For a diatomic molecule $I = \mu m r$. The reduced mass μ of a diatomic molecule with constituent atoms of mass m_1 and m_2 , is $\mu m = (m_1 m_2)/(m_1 + m_2)$. The equilibrium separation of the C and O atoms of CO is $r = 0.112$ nm.

Molecule	**Transition**	n_c	E/k_B (K)	ν (GHz)
^{12}CO	$J = 1 \rightarrow 0$	1.9×10^3	5.5	115.271
^{13}CO	$J = 1 \rightarrow 0$	1.7×10^3	5.3	110.201
C^{18}O	$J = 1 \rightarrow 0$	1.7×10^3	5.3	109.782
^{12}CO	$J = 2 \rightarrow 1$	6.3×10^3	11.1	230.538
^{13}CO	$J = 2 \rightarrow 1$	5.4×10^3	10.6	220.399
C^{18}O	$J = 2 \rightarrow 1$	5.5×10^3	10.5	219.560
^{12}CO	$J = 3 \rightarrow 2$	1.6×10^4	16.6	345.796
^{13}CO	$J = 3 \rightarrow 2$	1.4×10^4	15.8	330.588
C^{18}O	$J = 3 \rightarrow 2$	1.4×10^4	15.7	329.331

TABLE 1.2: The critical (number) densities n_c, excitation energies E/k_B, and frequencies (ν) of the lowest-lying (and most frequently observed) rotational transitions of the most common CO isotopologues (Draine, 2011).

were estimated by comparing the intensities of molecular lines in rare species or highly optically thin regions yielding to be X(^{12}CO/^{13}CO) $\approx$ 77 and X(^{12}CO/C^{18}O) $\approx$ 560 for conditions matching the Solar neighbourhood (Wilson & Rood, 1994). These relative abundances however do vary across the Galaxy, and there is evidence of a gradient in (^{12}CO/C^{18}O and ^{12}CO/C^{13}O) increasing from the Galactic centre outward(Langer & Penzias, 1990; Milam et al., 2005).

Smoothed particle hydrodynamics simulations of molecular clouds (Duarte-Cabral & Dobbs, 2016) have shown that CO emission traces density peaks of H_2 accurately, while it misses diffuse gas. Non-emitting CO is expected where H_2 densities are below the critical density. In cold ($T \leq 20$ K) and dense (10^5 particles per cm^3) environments CO can get trapped on the surfaces of dust grains. Depletion factors vary between 10 and 80 are typical in dense regions (Pon et al., 2016) in dense regions of IR dark clouds (Fontani et al., 2012).

Despite accounting for only $\sim$ 1% of the ISM, dust whose grains consist of tens or hundreds of atoms, is another important tracer of molecular gas (see section 1.1). Nearby ($<$ 500 pc) molecular clouds can be detected as optical absorption features against a background of starlight. For clouds at greater distances massive IR dense clouds (IRDC) have column density large enough to absorb in mid-IR (Peretto & Fuller, 2009). The H_2 column density can be determined from the level of absorption of dust by converting the reddening of the emission to the column density of atomic and molecular hydrogen (Bohlin et al., 1978; Fitzpatrick, 1999). The thermal emission of dust grains can be directly observed at far-IR, submillimetre, and millimetre wavelength. Under the assumption that a single temperature can be assigned to dust grains, it becomes possible

to estimate the H_2 column density averaged over a telescope beam (e.g. Schuller et al., 2009).

1.4 Emission segmentation

The study of molecular emission has been approached through a wide range of analytic methods. Each technique focuses on the analysis of a different feature of the gas. Structural patterns in molecular emission have been investigated through fractal analysis (Stutzki et al., 1998), the study of power spectra (Lazarian & Pogosyan, 2000) and the structure-function (Heyer & Brunt, 2004) have aimed to characterise turbulence in clouds (Brunt et al., 2010; Brunt & Federrath, 2014), and clump identification algorithms (Stutzki & Güsten, 1990; Berry, 2015; Colombo et al., 2015a) have been used to probe geometry, structure and substructure, e.g. the density hierarchy.

In general, statistical approaches to the analysis of molecular line data either aim to provide a statistical description of the emission over the entire dataset or a division of the emission into physically relevant features. The latter approach is then followed by the analysis of the characteristics of the resulting population of sources. Statistical analysis include fractal analysis (Elmegreen & Falgarone, 1996; Stutzki et al., 1998; Elmegreen, 2002; Sánchez et al., 2005; Lee et al., 2016), Δ-variance (Stutzki et al., 1998; Klessen & Glover, 2015), correlation functions (Houlahan, 1990; Rosolowsky et al., 1999; Lazarian & Pogosyan, 2000; Padoan et al., 2003) and analysis of the two-dimensional power spectrum (Schlegel & Finkbeiner, 1998; Pingel et al., 2018; Combes, 2012; Feddersen et al., 2019) and principal components (Heyer & Brunt, 2004). These techniques provide the overall statistical properties of the sample and are thus best suited for the comparison of measurements between different datasets. On the other hand, clump identification (image segmentation) is preferred for the study of physically important substructures embedded in the emission.

In position-position velocity (PPV) data sets, giant molecular clouds (GMCs) and their substructure are identified as discrete features (sets of connected voxels) with emission (brightness temperature or column densities) above a specified threshold (Scoville et al., 1987; Solomon et al., 1987). Molecular cloud recognition in PPV data sets is performed with a variety of automatic algorithms.

These methods are commonly designed to operate on large data sets and different levels of blending between structures. Three different strategies for the identification of molecular emission are frequently employed in the construction of GMC identification software packages:

- the iterative fitting and subtraction of a given model to the molecular emission (Stutzki & Güsten, 1990; Kramer et al., 1998),
- the friends-of-friends paradigm that connects pixels based on their and their neighbours' emission values (Williams et al., 1994; Rosolowsky & Leroy, 2006),
- and gravitational acceleration mapping methods[4].

These approaches identify single objects by assigning individual pixels to partitions of the data set, thus recasting GMC recognition as an image segmentation problem (Pal & Pal, 1993). Contouring in three-dimensional images is however a complex task. Complications arise from the difficult deblending of internal structures in crowded regions as the often unclear boundaries that separate star-forming clouds from the surrounding multi-phase ISM (as the often unclear boundaries that separate star-forming clouds from the surrounding multi-phase ISM, see Ballesteros-Paredes et al., 1999; Hartmann et al., 2001; Blitz et al., 2007). The efficacy of the different classes of GMC recognition algorithms are thus affected by survey specific biases arising from spatial and spectral resolution and the sensitivity in molecular-line observations of GMCs (Rosolowsky & Leroy, 2006; Pineda et al., 2009; Wong et al., 2011). Cloud recognition usually worsens in regions characterized by complex molecular environments and crowded velocity fields (as the Inner Milky Way), where resolution plays a crucial role in the identification of structure (Hughes et al., 2013). At low resolution, segmentation algorithms suffer from the blending of emission from unrelated clouds Colombo et al. (2014), while high resolutions cause cloud substructures to be identified as individual clouds [5].

Dendrograms can be considered as graphical abstractions of the hierarchical structure of nested isosurfaces in PPV data. A dendrogram represents a reduction of the structure down to its defining features, thus it allows for the representation of a large, complex

[4] `https://arxiv.org/abs/1603.05720`

[5] In particular, friends-of-friends methods are especially sensitive to resolution. In clumpy environments, the objects naturally selected by this type of algorithm have the scale of a few resolution elements (Rosolowsky & Leroy, 2006).

molecular line dataset as a simple model through which we can probe the hierarchical structure of the emission at different spatial scales. It is important to notice that this approach to dendrogram formalism to represent contour surfaces differs significantly from its common uses in statistical analysis (Ghazzali et al., 1999). In a statistical analysis context, dendrograms usually serve as an intuitive representation of the clustering of a statistical set.

The definition of emission dendrograms is introduced by Rosolowsky et al. (2008) (also see B.2). This particular definition is a specific application of the more general 'structure trees' proposed and analysed by Houlahan & Scalo (1992) in their study of the characteristics of two-dimensional images. In particular, in this formalism, a dendrogram is a model that encodes and emphasises the properties (such as volume or density) of the isosurfaces present in three-dimensional emission datacubes.

1.5 A note on segmentation algorithms

In this work, two emission segmentation algorithms are considered and their performance is compared over the CHIMPS survey. The Spectral Clustering for Interstellar Molecular Emission Segmentation (SCIMES) is based on the graph-theoretical analysis of the emission dendrograms mentioned above and translates emission segmentation into a clustering problem (see Colombo et al., 2015a). While the FellWalker (FW) technique is a form of the watershed algorithm (see section Appendix B) designed to partition multi-dimensional arrays of data values into regions, each associated with significant peaks (Berry, 2015). By design, both these methods segment the input data array into disjointed subsets of data points. Although partitioning strategies are widely used, there exist alternative approaches to the recognition (extraction) of emission structures. Different paradigms that allow for the overlap of emission clumps can prove beneficial in different circumstances (Stutzki & Güsten, 1990; Men'shchikov et al., 2012). For instance, the deblending of crowded overlapping sources is the cause of major uncertainties in line-of-sight projected, two-dimensional images. This situation is aggravated by the presence of filaments whose orientation impacts the projection significantly (Men'shchikov et al., 2010; Arzoumanian et al., 2011). In these cases, allowing for the overlapping of clumps in emission extraction becomes beneficial for the interpretation of the observations.

FIGURE 1.2: Molecular clouds in M51, the Whirpool galaxy. The distribution of hydrogen molecules, blue markers, is superimposed to a colour image of M51. The location of molecular hydrogen has been traced through ^{12}CO (1 - 0) emission, as measured in the PdBI Arcsecond Whirlpool Survey (PAWS) study using the millimetre telescopes of the Institut de Radioastronomie Millimetrique (IRAM). Credit: PAWS Team/IRAM/NASA HST/T. A. Rector, University of Alaska Anchorage

1.6 Star formation in the Milky Way

Despite the progress on the characterisation of molecular clouds and their structure, devising a quantitative model, empirical or theoretical, that predicts the efficiency of star-forming processes and their relation to the physical properties of the interstellar gas is an elusive task.

Empirical relations such as Schmidt-Kennicutt (Kennicutt, 1998) suggest that the star formation is solely regulated by the amount of gas that exceeds a certain density threshold (Gao & Solomon, 2004; Lada et al., 2012; Evans et al., 2014; Zhang, 2014). However, these simple scaling laws are constrained by the sample population size and break down over scales smaller than a few hundred pc, where the enclosed sample of molecular clouds decreases significantly (Kruijssen & Longmore, 2014). Power spectrum studies of giant molecular clouds maps in the Galactic disk have shown that the SFE and clump

formation efficiency (dense gas mass fraction, DGMF) vary significantly on the scales of individual clouds peaking at 10-30 pc (Eden et al., 2021). This variation in SFE declines at a (smoothing) scale of 100 pc. Furthermore, it was found that the distributions of SFE and DGMF in individual clouds are consistent with being lognormal (Eden et al., 2012, 2013) and thus possibly a combination of several random factors implying that extreme star-forming regions (or regions in which star formation is absent) are not necessarily due to special conditions. These results are also consistent with a simple Schmidt-Kennicutt law since the distribution of SFEs possesses a well-defined mean when averaged over kpc scales and a large number of clouds. Furthermore, the SFE/DGMF appears to vary several orders of magnitude from cloud to cloud. Along with the nearly constant mean value of the distribution of SFEs, this fact suggests that differences between the individual clouds are more relevant to star formation than large-scale mechanisms such as density features, shear, and radial variations in metallicity. In particular, spiral arms appear to mainly only produce source crowding (Figure 1.2). Ragan et al. (2016) and Ragan et al. (2018) also confirmed no arm-associated signal in the fraction in the Hi-GAL catalogue of compact sources that are currently star-forming. These results agree with observations of spiral galaxies indicating that the H_2/HI fraction and the SFE traced by infrared (IR) and ultraviolet (UV) emission in spiral arms are not significantly higher than in the inter-arm gas (Kennicutt et al., 2003; Gil de Paz et al., 2007; Walter et al., 2008; Leroy et al., 2009; Obreschkow & Rawlings, 2009; Foyle et al., 2010). Also, the fraction of GMCs formed from HI appear to be determined by the H_2 formation/destruction rate balance and stellar feedback (Leroy et al., 2010). These mechanisms act at small scales in the ISM. Except for starburst galaxies and ultraluminous IR galaxies (ULIRGs), internal radiative feedback is expected to determine the properties of molecular clouds with the minor influence of the external environment (Krumholz et al., 2009). These pieces of evidence challenge the idea that spiral arms may be direct triggers of star formation. However, HI and CO data in W3, W4 and W5 showed that the molecular fraction of the gas content in the outer Perseus spiral arm is 10 fold higher than in the inter-arm regions (Heyer & Terebey, 1998), implying that spiral density waves both raise the efficiency with which molecular clouds are formed

and, consequently, the SFE in those regions (Dobbs et al., 2006)[6]. This is an as-yet unresolved contradictory piece of evidence involving molecular gas in the outer Galaxy (see section 8.1).

A key element for any predictive quantitative model (theoretical or empirical) for star formation is the mechanism regulating the Initial Mass Function (IMF) and its relation to SFE. The IMF cannot be observed directly but is modelled and the outcomes can be compared with observations. Following the pioneering work of Salpeter (Salpeter, 1955) there have been many studies of the IMF in various regions of the Milky Way and other galaxies producing standard forms of the IMF such as the Miller-Scalo, Kroupa, and Chabrier (see Kroupa et al., 2013, for a review). There is no reason to believe that the IMF should be universal[7] and although systematic variations in the IMF depending on environmental conditions have been surprisingly small (Bastian et al., 2010), there is mounting observational and theoretical evidence that challenges the IMF universality. These studies include observations of H_α and the far-UV emission of HI in external galaxies Meurer et al. (2009), optical observation of ultra-faint satellites of the Milky way (Gennaro et al., 2018), H_β imaging of gas-rich, star-forming nearby dwarf galaxies, the investigation of mass segregation in starburst clusters (Dib et al., 2007; Dib, 2014), magneto-hydrodynamical simulations (Ferré-Mateu et al., 2013) and Montecarlo simulations (Dib et al., 2017). Variations in the IMF affect the estimation of stellar masses from photometry and the gas mass budget between different generations of stars, resulting in a modified characteristic formation timescale of galaxies. In this framework, the IMF could also mimic SFE changes as measured by the L/M parameter (see Chapter 6). Observation of the R136 star cluster in the Galactic Centre support this idea (Crowher et al., 2010).

The problem of setting up a comprehensive model for SFE is further aggravated by the impact of large-scale radial changes in Galactic environments on the star-forming properties of the gas. The fraction of molecular gas has been observed to decrease

[6]Other studies Dobbs et al. (2011) interpret spiral arms as organising features that affect the ISM by delaying and crowding the gas, which is deflected from circular orbits when it enters the arm. The star formation rate in the arm is thus increased by enabling longer-lived and more massive molecular clouds. In this framework, molecular clouds with spiral arms have longer lifetimes than those in the inter-arm gas, resulting in longer star formation time scales and consequently an increased SFE (Roman-Duval et al., 2010).

[7]A universal IMF may be produced by a universal physical process, but the converse is not true: a universal physical process does not necessarily lead to a universal IMF (Narayanan & Davé, 2012; Hopkins, 2013).

rapidly with Galactocentric distance, from ≈ 100% within 1 kpc to only a few per cent at radii greater than 10 kpc (Sofue & Nakanishi, 2016). Simultaneously, DGMFs peak at around 3–4 kpc and then decline in the inner zone, where the disc becomes stable against gravitational collapse on large scales. This is the zone swept by the Galactic bar and star formation is suppressed for the life of the bar (James & Percival, 2016). The SFE, measured as either the integrated infrared luminosity from young stellar objects (YSOs) or the numbers of HII regions per unit molecular gas mass, is low but steady on kiloparsec scales at radii greater than 3 kpc. The SFE declines abruptly in the Central Molecular Zone (CMZ) within 0.5 kpc (Longmore et al., 2013; Urquhart et al., 2013). This significant difference may be related to higher turbulent gas pressure in the CMZ, which raises the density threshold for star formation (Kruijssen & Longmore, 2014), but the cause of such differences and transitions between these regions remains unexplained. The low SFE in the CMZ cloud G0.253+0.016 appears to be caused by a prevalence of shear-driven solenoidal (divergence-free) turbulence modes, in contrast to spiral-arm clouds, which typically have a significant compressive (curl-free) component (Federrath et al., 2016). A similar analysis of the Orion B molecular cloud (Orkisz et al., 2017) finds that the turbulence is mostly solenoidal, consistent with its low SFR, but is position-dependent within the cloud, motions around the main star-forming regions being strongly compressive. Thus, this significant inter-cloud variability of the compressive/solenoidal mode fractions may be a decisive agent of variations in the SFE. The SFE may also be affected by cloud collisions, which should produce highly compressive gas flows.

The ^{13}CO/C^{18}O ($J = 3 \rightarrow 2$) Heterodyne Inner Milky Way Plane Survey (CHIMPS, Rigby et al. (2016)) has produced a large sample of molecular clouds and the first large-scale map of molecular-gas temperatures. This survey has also led to the discovery of significant new arm structures. Contrary to theoretical predictions (Kruijssen & Longmore, 2014), the study of CHIMPS clouds (Rigby et al., 2019) revealed SFE is neither linked to turbulent pressure nor Mach numbers in the disc.

Together, these findings emphasise the need for the detailed analysis of large samples of molecular clouds from different regions in the galaxy, relating their internal and external environmental conditions to their SFE and DGMF, as the next step in understanding the physics of star formation.

1.7 Goals and structure of the thesis

The aim of this work is two-fold. First a comparison between two catalogues of sources over CHIMPS and some of their characteristic physical properties is presented. One catalogue is constructed with the well-established watershed algorithm FellWalker (an extraction particularly popular among the users of the Starlink JCMT software suite), while for the other the more recent approach to cloud segmentation through spectral clustering, SCIMES, is employed. The second part of the project is a full sample study of turbulent modes in CHIMPS molecular clouds with a focus on their relation to star formation efficiency. The sample employed is a sub-catalogue of the SCIMES segmentation introduced above. The thesis is organised as follows. Chapter 2.1 is a brief introduction to the four CO surveys (CHIMPS, COHRS, ATLASGAL, and Hi-GAL) whose data are used both in the determination of the distances and physical properties in the SCIMES catalogue and the definition of a measure for star formation efficiency. Source extraction methods are mentioned to emphasise the variety of methods used and the need for systematic comparison of their effects of the emission extracted. For ATLASGAL we also describe the distance assignment method as this catalogue is the main reference for distance in CHIMPS. Chapter 3 describes the methods used for the preparation and post-processing of the CHIMPS for the construction of the SCIMES catalogue. Vital to the estimation of the physical quantities included in the catalogue is the algorithm used for distance assignments to the sources identified through SCIMES. The Chapter includes the FellWaker and SCIMES algorithms as well. Chapter 4 describes the construction of a SCIMES catalogue of the emission in CHIMPS and compares the characteristic to the FW extraction. Chapter 5 introduces the statistical method devised by Brunt et al. (2010); Brunt & Federrath (2014) which allows for the analysis of turbulent modes within molecular clouds from a line-of-sight projected data set. This Chapter is followed by an analysis of the star formation efficiency over a full sample selection of clouds from the SCIMES catalogue. Chapter 7 summarises the results found in the thesis and in Chapter 8 plan for the continuation and extension of the analysis that was initiated with the present thesis is proposed.

Finally, Appendix A consists of a short description of the FellWalker watershed algorithm. Appendix B presents a detailed description of the SCIMES algorithm including an introductory explanation of the graph-theoretical concepts upon which the algorithm

is based. Appendix C revisits the derivation of Brunt's method in full detail. Appendix D concerns random distance assignment and the resulting mass distributions. Appendix E describes the FINDBACK algorithm used to implement noise removal through a multi-step smoothing filter. Appendix F collects the graphical representations of the FW distance assignments within SCIMES clouds over the ten regions covered by CHIMPS.

Chapter 2

Surveys

This Chapter provides an overview of the surveys used for the analyses presented in this thesis. At the core of these studies lie the CO emission data collected in the CO Heterodyne Inner Milky Way Plane Survey (CHIMPS). Most of the Chapter is thus dedicated to the description of CHIMPS and the derivation of column density and excitation temperature maps for this dataset. Other surveys such as COHRS, Hi-GAL, ATLASGAL were used in the distance assignments and the determination of star formation efficiency, along with the comparison of the distribution of the physical quantities associated with molecular clouds. A brief description of these surveys is also included. The sections dedicated to each survey include subsections with a focus on the aspects of the data that are required for the analysis presented in this thesis (e.g. distance assignments in ATLASGAL, see also Chapter 4). Mentions of the emission extraction algorithms are also made to emphasise the diversity of the method employed in different projects.

The CHIMPS (FellWalker extraction) and COHRS catalogues used in the analysis in Chapter 4 were produced, published, and made available publicly by their respective authors (Colombo et al., 2019; Rigby et al., 2019). The luminosities and bolometric temperatures used to derive the star formation efficiency in Chapter 6 were taken as published by Molinari et al. (2016).

2.1 CHIMPS

The ^{13}CO/C^{18}O (3−2) Heterodyne Inner Milky Way Plane Survey (CHIMPS) is a spectral survey of the $J = 3 - 2$ rotational transitions of ^{13}CO at 330.587 GHz and C^{18}O at 329.331 GHz. The survey covers ∼19 square degrees of the Galactic plane, spanning longitudes l between 27.5° and 46.4° and latitudes $|\,b\,| < 0.5°$, with angular resolution of 15 arcsec. The observations were made over a period of 8 semesters (beginning in spring 2010) at the 15-m James Clerk Maxwell Telescope (JCMT) on Manua Kea in Hawaii. Both isotopologues were observed concurrently (Buckle et al., 2009)) using the Heterodyne Array Receiver Programme (HARP) together with the Auto-Correlation Spectral Imaging System (ACSIS). The HARP array is composed of 16 (4×4) focal plane superconductor–insulator–superconductor heterodyne detectors. The spacing between consecutive receptors corresponds to 30 arcsec on the sky. HARP operates at submillimetre frequencies between 325 and 375 GHz. ACSIS was set to a total bandwidth of 250 MHz, 61 kHz for each of its 4096 frequency channels. With a velocity width of 0.055 $\mathrm{km\,s^{-1}}$ per channel, CHIMPS spans a velocity bandwidth of $\sim 200\ \mathrm{km\,s^{-1}}$. The data are structured as position-position-velocity (PPV) cubes with velocities binned in 0.5 $\mathrm{km\,s^{-1}}$ channels and a bandwidth of 200 $\mathrm{km\,s^{-1}}$. The Galactic velocity gradient associated with the differential rotation of the Galaxy is matched by shifting the velocity range with increasing Galactic longitude from $-50 < v < 150\ \mathrm{km\,s^{-1}}$ at 28° to $-75 < v < 125\ \mathrm{km\,s^{-1}}$ at 46°.

2.1.1 Observations and data

The observation mode consists of a position-switched raster. This mode scans the sky with a chosen width in a pattern that fills the image pixels from edge to edge and back, from bottom to top. At the end of each row, the receptor array is shifted by half its width perpendicularly to the scanning direction. Each point in the observation area is thus scanned by several detectors. Then, a second scan is performed, repeating the same pattern, but perpendicular to the first pass. Off-positions are taken below the Galactic plane with a latitude offset of $\Delta = -1.5°$ for each scan. This observation mode produces a sample spacing of 7.3 arcsec and a sample time of 0.25 seconds, yielding approximately a 21×21 arcminute datacube per hour. The pointing accuracy at JCMT

is approximately 2 arcsec in azimuth and elevation (checked between observations). The JCMT tracking is generally more accurate than 1 arcsec for each hour of observation.

Raw data are recorded continuously during the scans in a time-series format. Calibrations of the spectra occur during the observations (Kutner & Ulich, 1981, three-load chopper-wheel method). The intensity of the emission is recorded as the corrected antenna temperature, T_A^*, a temperature scale that accounts for atmospheric attenuation, ohmic losses inside the instrument, spillover, and rearward scattering (Rigby et al., 2016). The T_A^* scale is calibrated absolutely against spectral standards observed and updated nightly[1]. The tolerance for integrated intensities and calibrated peak emission is 20% of the values of the standards. Receivers with readings of the standards with absolute values that exceed this tolerance are re-tuned. The main beam brightness temperature can directly be recovered from the corrected antenna temperature:

$$T_{\rm mb} = \frac{T_A^*}{\eta_{\rm mb}}, \tag{2.1}$$

where $\eta_{\rm mb} = 0.72$ is the mean detector efficiency (Buckle et al., 2009).

To convert the raw time-series spectra to spectral data cubes with an associated coordinate grid, the standard JCMT ORAC-DR data reduction pipeline (Jenness et al., 2014) that employs the KAPPA, SMURF, and CUPID packages included in the Starlink (Currie, 2013) suite, was used. In particular, the narrow-line reduction was applied (Cavanagh et al., 2008). This reduction routine is specifically optimised for the reduction of narrow-line-width and low-velocity-gradient sources. The process has two main stages: the quality assurance of the data and the iterative construction of the spectral cubes and other outputs.

In the reduced cubes, the pixel size is set to 7.6 arcseconds in Galactic longitude and latitude, while the velocity channels are 0.5 $\mathrm{km\,s^{-1}}$ to improve the signal-to-noise ratio (SNR). The observations' raster pattern results in reduced cubes that are under-sampled at the edges (where the scanning array changes direction). These areas also present a lower SNR. The data values at the edges are adjusted by cropping the cubes. Cropping produces overlapping (approximately 1 arcmin in width) between adjacent cubes.

[1] `http://www.eaobservatory.org/jcmt/instrumenation/heterodyne/calibration`

The reduced data cubes each include a variance array component. The ^{13}CO survey has mean rms sensitivities of $\sigma(T_A^*) \approx 0.6\,\mathrm{K}$ per velocity channel, while for C^{18}O, $\sigma(T_A^*) \approx 0.7\,\mathrm{K}$. These values, however, fluctuate across the survey region depending on both weather conditions and the varying numbers of working receptors on HARP (Rigby et al., 2016). The rms of individual cubes range between 0.37 K and 1.51 K and between 0.43 K and 1.77 K per channel for the for the ^{13}CO and the C^{18}O emission respectively.

The reduced data are organised into 178 datacubes which are, in turn, mosaiced into 10 larger regions (see also Chapter 3, Figure 3.5) since the entire CHIMPS area is too large to be analysed as a single datacube. Each of the regions contains a variance array component determined for each spectrum from the system noise temperature. In order to perform source extraction as consistently as possible, a small overlap is left between adjacent regions. Both the regions and the cubes that constitute them are available for download in FITS format from the Canadian Archive Network for Astronomical Research (CANFAR)[2]. The data are presented in corrected antenna temperature in units of K. Column density and excitation temperature maps for the 10 regions can also be obtained from the CANFAR servers.

2.1.2 Column density and excitation temperature maps

The total column density throughout a CHIMPS datacube can be calculated from the excitation temperature and the optical depth of the CO emission. This calculation is outlined in Rigby et al. (2019). Their method is a variation of the standard approach for the determination of the excitation temperature and optical depth and uses ^{13}CO$(3-2)$ emission at each position (l, b, v) in the datacube (on a voxel-by-voxel basis) under the assumption of local thermal equilibrium (Roman-Duval et al., 2010). This strategy has a major advantage over the analysis of velocity-integrated properties: any property derived from the excitation temperature and optical depth is independent of the source extraction and image segmentation algorithms. However, an analysis based on individual voxel information does not account for the attenuation of the emission due to self-absorption. Although, Rigby et al. (2019) performed the first-order adjustment of their method with respect to the ^{12}CO$(3-2)$ from which excitation temperature of ^{13}CO$(3-2)$

[2]https://doi.org/10.11570/19.0028

is derived, they did not find evidence for significant self-absorption in ^{13}CO$(3-2)$ across the entire CHIMPS area.

The total column density at each position, $N_{13}^{\rm Tot}$, is determined from the column density, $N_{13}(J)$, within a specific energy level, J, by multiplying it by a partition function representing the sum over all states, giving

$$N_{13}^{\rm Tot} = N_{13}(J)\frac{Z}{2J+1}\exp\left(\frac{hBJ(J+1)}{k_B T_{T_{\rm ex}}}\right), \tag{2.2}$$

where h is the Planck constant, k_B is the Boltzmann's constant, $T_{\rm ex}$ is the excitation temperature, and $B = h/(8\pi^2 I)$ with the moment of inertia $I = \mu R_{\rm CO}^2$ calculated from the reduced mass μ and the mean atomic separation $R_{\rm CO} = 12$ nm. Assuming that the vibrationally excited states are not populated, Z can be approximated as

$$Z \approx \frac{k_B}{hB}\left(T_{\rm ex} + \frac{hB}{3k_B}\right). \tag{2.3}$$

Within the $J = 2$ state, the column density $N_{13}(J = 2)$ (number of CO molecules per cm^2) is calculated as

$$N_{13}(J=2) = \frac{8\pi}{c^3}\frac{g_2\nu^3}{g_3 A_{32}}\frac{1}{1-\exp(-h\nu/k_{\rm B}T_{\rm ex})}\int \tau_v dv, \tag{2.4}$$

with g_2 and g_3 being the statistical weights of the $J = 2$ and $J = 3$ rotational energy levels respectively. The constant $A_{32} = 2.181 \times 10^{-6}$ s^{-1} is the Einstein coefficient for the ^{13}CO$(3-2)$ transition (Schöier et al., 2005), τ_v is the optical length at the frequency v the frequency v in GHz, and the velocity channel width dv is given units of km s^{-1} . There is a small discrepancy between the values of Z as defined in equation 2.3, and those reported in the Cologne Database for Molecular Spectroscopy 2 (Endres et al., 2016). This difference is due to the hyperfine splitting of ^{13}CO$(3-2)$, which is not accounted for in equation 2.3. The impact on the column densities consists of a variation of $0.5-2\%$ over a temperature range of $5-20$ K. These discrepancies are not significant for the purpose of our investigation.

2.2 COHRS

The JCMT ^{12}CO $(3-2)$ High Resolution Survey (COHRS) is a large-scale CO survey that mapped the ^{12}CO $(3-2)$ emission in the Inner Milky Way plane. The survey covers latitudes $10.25° < l < 17.5°$ with longitudes $|b| \leq 0.25°$ and $17.5° < l < 50.25°$ with $|b| \leq 0.25°$. This particular region was selected to match a set of important surveys, among which CHIMPS, the Galactic Ring Survey (GRS, Jackson et al., 2006), the FOREST Unbiased Galactic plane Imaging survey with the Nobeyama 45-m telescope survey (FUGIN, Umemoto et al., 2017, see Figure 2.1), the Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE, Churchwell et al., 2009a), the Bolocam Galactic Plane Survey (BGPS, Aguerre et al., 2011), and the Herschel Infrared Galactic Plane Survey (Hi-GAL, Molinari et al., 2016). The observations were performed with the Heterodyne Array Receiver Programme B-band (HARP-B) at 345.786 GHz and ACSIS set at a 1 GHz bandwidth yielding a frequency resolution of 0.488 MHz (0.42 $\mathrm{km\,s^{-1}}$). The survey covers a velocity range between -30 and 155 $\mathrm{km\,s^{-1}}$ with a spectral resolution of 1 $\mathrm{km\,s^{-1}}$ and angular resolution of 16.6 (FWHM). The COHRS data (first release) are publicly available[3].

2.2.1 Catalogue

Molecular clouds in the reduced COHRS data were identified through the SCIMES method by Colombo et al. (2019). Before the SCIMES algorithm is applied, the tiles are mosaicked together into a single survey-wide cube. To highlight emission features by increasing the SNR the data were masked multiple times before being divided again into smaller regions. Pre-segmentation masking and the construction of the final datacubes are explained in Colombo et al. (2019) and Rosolowsky & Leroy (2006). The final cubes span 1200 pixels in longitude corresponding to $\sim 2°$ in longitude. The following parametrisation of SCIMES is run on each region, all of the emission in the mask is considered (`min_val` $= 0$). Each dendrogram branch should have an intensity change greater than $3\sigma_{\mathrm{rms}}$, `min_delta` $= 3\sigma_{\mathrm{rms}}$ and contain at least as many pixels as three resolution elements (`min_val` $= 3\Omega_{\mathrm{bm}}$, where Ω_{bm} is the solid angle subtended by the beam expressed in pixels.

[3] `http://dx.doi.org/10.11570/13.0002`

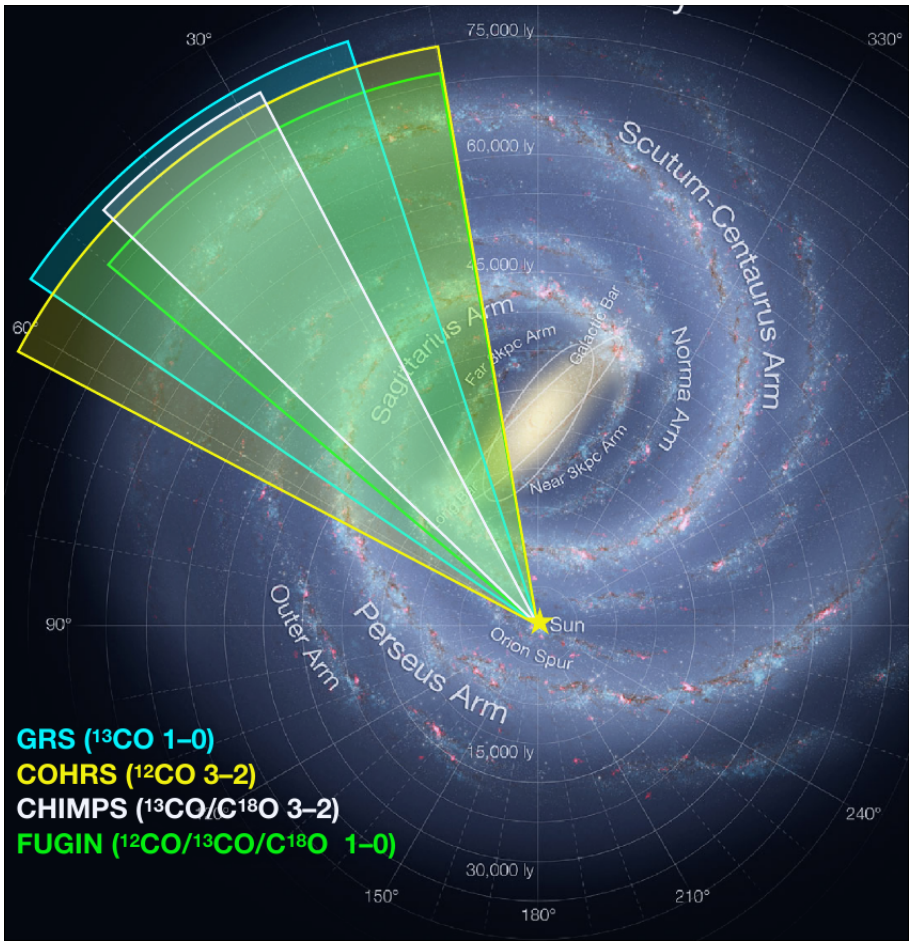


FIGURE 2.1: Coverage of the Galactic plane by four main CO surveys: CHIMPS, COHRS, the Galactic Ring Survey (GRS, Jackson et al., 2006) and the FOREST Unbiased Galactic plane Imaging survey with the Nobeyama 45-m telescope survey (FUGIN, Umemoto et al., 2017).

It is important to notice that this parametrization was specifically chosen for the segmentation of ^{12}CO emission. Its results cannot directly be compared to the emission features found through the extraction of a different isotopologue with its own SCIMES parametrization. However, as it will be shown in the next Chapter, the information from different segmentations can be used to complement each other. The most compact structures/star formation sites identified in the ^{13}CO emission can be matched to the $J = 3 - 2$ transition of the ^{12}CO isotopologue that traces warm molecular gas (10-50 K) around the active star formation regions. The volume and luminosity affinity matrices (see Appendix B) are constructed using the PPV volumes and integrated intensity values since spatial volumes and intrinsic luminosities cannot be used without knowing the distances to the dendrogram branches. The scaling parameter is set above $3\sigma_{\rm rms}$ (Colombo et al., 2015b) .

2.2.2 Distances

Distance assignments follow Zetterlund et al. (2018). Their estimation of distances is based on an analysis of the BGPS (Aguerre et al., 2011) along with Reid's kinematic distance calculator (Reid et al., 2016). Column densities are calculated by scaling the integrated intensities of the CO emission with a H_2-to-CO conversion factor. Masses are estimated directly through the distances of the molecular clouds. The COHRS catalogue also includes an additional dynamical measurement of mass, the virial mass. However, this calculation requires the assumption of virialised spherical clouds with a density profile that decays as r^{-1}. External pressure and magnetic fields are also assumed to be negligible. Detailed calculations of both the pixel-based and physical properties of the COHRS molecular clouds (in particular the effective radius, velocity dispersion, and CO luminosity from which all other properties are derived) are presented in Colombo et al. (2019). The COHRS cloud catalogue is available on the publisher's site [4].

2.3 ATLASGAL

The Atacama Pathfinder Experiment (APEX) is a 12 m single-dish submillimetre telescope located on the Chanjnantor Plateau in Chile. It is equipped with heterodyne receivers with frequencies ranging from 230 GHz to 1.4 THz and several arrays of bolometers. The APEX Telescope Large Area Survey of the Galaxy (ATLASGAL) is an unbiased survey that observed the Galactic plane with the Large APEX BOlometer CAmera (LABOCA) at 870 μm. LABOCA consists of 295 bolometer detectors arranged in a hexagonal pattern yielding a field of view of 11.4′ in diameter. ATLASGAL covers Galactic longitudes $60° < l < 300°$ and latitudes $|b| < 1°.5$ and is one of the largest and most sensitive ground-based submillimetre Galactic surveys. ATLASGAL is believed to detect all dense clumps with mass $> 1000 M_\odot$) with heliocentric distance < 20 kpc in the Milky Way and to encompass samples representing all stages of high-mass star formation. The ATLASGAL survey has been the basis for many studies of the distribution of Galactic dense molecular gas (Beuther et al., 2012; Csengeri et al., 2014) and the ATLASGAL Compact Source Catalogue (CSC, Contreras et al. (2013); Urquhart et al.

[4] https://doi.org/10.1093/mnras/sty3283

(2014a)) is a comprehensive catalogue of over 10000 dense clumps extracted from the reduced ATLASGAL data.

2.3.1 Data

The raw data are recorded in MB-FITS (Multi-Beam FITS) format by the APEX Control System (APECS, Muders et al., 2006). The BOlometer array data Analysis package (BOA, Schuller, 2012), an algorithm specifically optimized for reduction of LABOCA data was employed in the pipeline.

The Source Extractor software (SExtractor, Bertin & Arnouts, 1996) was used to segment the data images. SExtractor performs a complete analysis and extraction of an image in the following steps. To be able to detect the faintest emission, SEXtractor first runs a background estimator to construct a map of the background sky. This is accomplished by applying the estimator to each pixel of the image to determine the noise level. With a background estimate, source detection occurs via thresholding (masking out emission below a given threshold). Source deblending is the next stage of the process. Deblending separates adjacent objects that have been identified as a single source. Deblending in SExtractor is implemented as a multiple-isophotal technique. Each extracted source is re-thresholded at 30 levels, exponentially spaced between its primary extraction value and its peak values. This produces a dendrogram of the emission distribution (see section B.2), which is scanned from top to bottom (highest branch to the trunk) to check for source separation at the junctions of the branches. If the integrated pixel intensity of a branch is above a given fraction of the total intensity of the composite object, the branch is considered a separate source. Notice that this condition has to hold for at least two branches at the same emission level. Spurious sources, resulting, for instance, from low thresholds[5] are filtered out. This cleaning process considers the contribution to the mean surface brightness of each extracted source from its neighbours. This value is then subtracted from the source and its new emission is checked against the detection threshold. To reduce the occurrence of spurious sources and avoid missing real emission features, the reduced emission maps were converted to SNR maps (Contreras et al., 2013).

[5] A local higher background causes a lower relative threshold, which in turn leads to the detection of more noise peaks.

In the ATLASGAL CSC, SExtractor was used to calculate the positions of peaks, fluxes, and the size of the sources. As the catalogue was meant to solely contain compact sources, a threshold (4) was put on the ratio of semimajor to semiminor axis of the ellipse approximating the source. A detailed description of the catalogue is given in Contreras et al. (2013).

2.3.2 Radial velocities

To determine the distance and the physical properties associated with a source, its radial velocity with respect to the local standard of rest (LSR) is required. Together with a model of the Galactic rotation curve, radial velocities make it possible to determine kinematic distances. Urquhart et al. (2018) elaborated the original ATLASGAL CSC to include radial velocities and distance estimates. The radial velocities of the clumps were measured from molecular line observations (in particular CO, NH_3 and CS). Molecular line measurements that match most of the ATLASGAL CSC entries are found in a variety of surveys. In particular, the following Galactic plane surveys were used: Galactic Ring Survey (GRS, Jackson et al., 2006), Mopra CO Galactic plane Survey (MGPS, Burton et al., 2013), the Three- mm Ultimate Mopra Milky Way Survey (ThrUMMS, Barnes et al., 2015), (SEDIGISM, Schuller et al., 2017), COHRS (Dempsey et al., 2013), CHIMPS (Rigby et al., 2016) in combination with selected samples from large observational programs: The Millimetre Astronomy Legacy Team 90 GHz Survey (MALT90, Jackson et al., 2013)), the Red MSX Source survey (RMS, Urquhart et al., 2007, 2008, 2011, 2014b), BGPS (Aguerre et al., 2011), dedicated ATLASGAL follow-up observations (Wienen et al., 2012; T. et al., 2016; Kim et al., 2017). To assign radial velocities counterparts of ATLASGAL clumps were searched for in molecular line catalogues. A velocity value is assigned to an ATLASGAL CSC source when the pointing centre of the molecular line observation is found to lie within the area of the source. When a spectrum at the source position contains more than one emission line the transition with the highest critical density is chosen. This means that, NH_3 and HNC are preferred to CO. Higher critical density means that the emission is less affected by multiple components that originate from the diffuse gas along with the sight between the source and the observer. The spectra of sources that lacked a known counterpart in the surveys were extracted from survey datacubes (after reduction and calibration). A Gaussian profile was used to fit these spectra. Unreliable fits (due to the data contamination by external

emission or strong baseline ripples) were discarded. Velocities were assigned as follows, for single component detections, the peak velocity of the molecular line was applied to the source. In the case of detections of multiple components with the strongest components within 10 km s^{-1}, the velocity of the strongest component was assigned. For other multiple component detections, the velocity of the component with the largest integrated line intensity was chosen (if the second strongest component had integrated line emission half as large). In all other cases, no velocity was chosen, additional observations were obtained and the allocation process was repeated.

2.3.3 Distances

From the radial velocity values in the ATLASGAL CSC, it is possible to estimate the heliocentric distances to the sources through the calculation of kinematic distances. Obtaining an accurate distance measurement is crucial for the calculation of many physical properties associated with a source. As mentioned above, the estimation of kinematic distances requires a model of the rotation of the Milky Way. A number of models describing the Galactic rotation curve have been developed during the years (Clemens, 1985; Brand & Blitz, 1993; Reid et al., 2014). All of them however yield kinematic distances that agree within their associated uncertainties (typically $\pm 0.3 - 1$ kpc) making them basically interchangeable. For the ATLASGAL CSC, the rotation curve devised by Reid et al. (2014) was adopted. This particular model is constrained by maser parallax measurements ($\sim$ 150 distances) and is known to produce kinematic distances that are comparable to maser distances [6].

Ambiguities arise intrinsically in the determination of kinematic distances. For the set of sources within the Solar circle, there are, in fact, two separate distance solutions that correspond to the same radial velocity. These distances are equispaced on both the 'near' and 'far' side of the tangent point [7]. To resolve this kinematic distance ambiguity

[6] Parallax measurements of masers observed in star-forming regions provide the most accurate distance assignments for molecular clouds. These compact bright sources in the ISM are powered by the emission (population inversion, rotational transitions, and collision (Gray et al., 2016)) from molecules such as water, hydroxyl radicals, methanol, formaldehyde, and silicon monoxide. Since masers are not associated with all sources (especially in the earliest stages of stellar evolution), however accurate, this method cannot be applied globally. Also, the difficulties related to maser parallax measurements and the low coverage of such sources in the Southern hemisphere (Reid et al., 2014) result in a limited number of distances known in the literature, the vast majority of which are located in the first two Galactic quadrants.

[7] With reference to Figure 2.2 (see Figure 2.2), the distance to the source in terms of the Galactic longitude l and the distance of the sun to the Galactic centre, R_0, is shown to be

(KDA) and make unique distance assignment Reid et al. (2016) developed a Bayesian maximum likelihood method. Under the assumption that each source is likely to be found within spiral arms, their method returns a unique distance that accounts for the relative positions of the spiral arms, the latitude of the source, and a probability to find the source at the near/far distance. Urquhart et al. (2018) applied both Reid's rotation curve and Bayesian maximum likelihood models to the ATLASGAL sources finding that the letter gave more reliable distances for sources located near the Solar circle. The overall difference in distance between the two methods is relatively small, amounting to less than 1 kpc in $\sim$ 95% of the sources. In particular, this difference becomes negligible in the fourth quadrant where the lack of maser parallax distances does not allow for an accurate model of the spiral arms' positions.

To provide distance assignments that overcome the KDA and to avoid the binding of sources to spiral arms[8], Urquhart et al. (2018) addresses the KDA with a series of alternated checks based on the known information on the source and its environment. Their scheme is reproduced in the flowchart in Figure 2.3. When possible, the sources within the Solar circle ($r < 8.35$ kpc) are matched to clumps with reliable distances from the literature: maser parallax, (Reid et al., 2014) and spectroscopic measurements (Moisés et al., 2011). If a known distance is found, it is assigned to the source. All sources with velocities close to the tangent velocity $|v_{\rm source} - v_{\rm tan}| < 10$ km s^{-1} are simply assigned the tangent distances since the difference between their far and near distances are smaller than their uncertainties.

Studies of high-mass stars within the Solar circle (Reed, 2000; Green & McClure-Griffiths, 2011; Urquhart et al., 2014b) have shown that their latitude distribution is correlated with the Galactic mid-plane. Assuming a similar distribution of high-mass star-forming clamps, the distances of clumps located within the Solar circle can be constrained in terms of the scale-height (Urquhart et al., 2018). If all sources from the Galactic mid-plane are initially assumed to be at the far distance, any with height above the mid-Plane of greater than 120 pc (four times the scale height) will be considered

$$d = R_0 \cos(l) \pm \sqrt{R^2 - R_0^2 \sin^2(l)}, \tag{2.5}$$

this solution gives rise to the KDA.

[8]To study how different Galactic environments and in particular the presence of spiral arms impact the star formation processes, inter-arm sources need to be separated from arm sources as much as possible. Reid's Bayesian maximum likelihood method promotes the allocation of sources to arms.

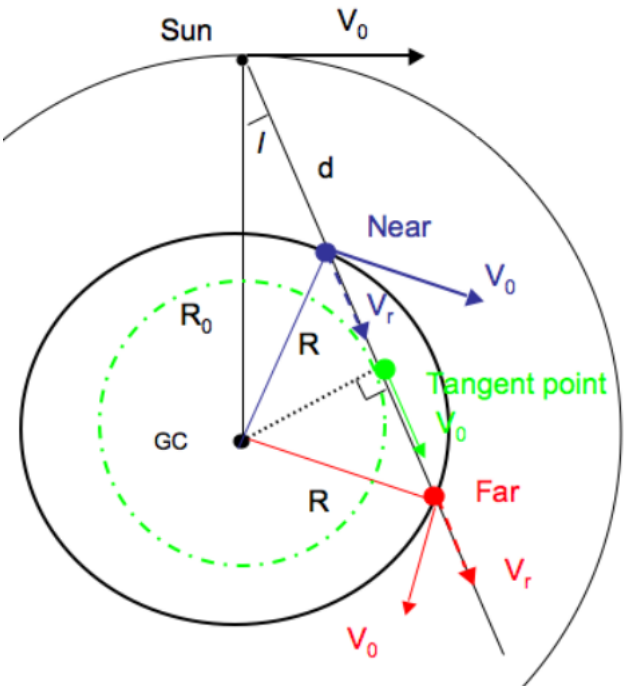


FIGURE 2.2: A geometric representation of the kinematic distance ambiguity for sources located within the Solar circle (Reid et al., 2014). GC denotes the Galactic centre

not reliable. In this case, the source's near distance is assigned (Urquhart et al., 2014b, 2018).

Distances to molecular clouds can also be determined through the analysis of the absorption features of the HI gas surrounding HII regions. The presence of HI along the line of sight manifests against the strong HII radio continuum by producing an adsorption feature at the velocity of the HI envelope (Wienen et al., 2015; Urquhart et al., 2012; Anderson, 2009; Kolpak et al., 2003). If the HII clumps are positioned at the near distance, absorption will be observed at the same velocity as the HII region, but not at higher velocities. Whereas, with HII at the far distance, absorption is expected at higher velocities than the HII emission source (extending all the way to the source's tangent velocity). The ATLASGAL sources were matched against clumps with HII region studies and assigned the distances found in the literature. If no such distance is known, HI absorption was checked (Urquhart et al., 2018). Clumps at the near distance are likely to absorb the emission from warm HI gas behind them. This results in an absorption feature in the HI spectra at the same velocity as the clump. This feature is absent when the clump is at the far distance with the warm HI gas being distributed throughout the

Galactic plane (Roman-Duval et al., 2010; Anderson, 2009; Jackson et al., 2006).

Finally, if the HI analysis is inconclusive, extinction towards the clumps is checked (Peretto & Fuller, 2009). If a clump is associated with an infrared dark cloud (IRDC) (Rathborne et al., 2006), the source is likely to be in the foreground with respect to the bright IR emission that present in the Inner Galaxy and its near distance is chosen.

Following the method shown in the flow diagram in Figure 2.3, Urquhart et al. (2018) assigned distance to $\sim$ 90% of the sources in the ATLASGAL CSC. The remaining clumps are either those lacking a radial velocity or those for which the distance ambiguity could not be resolved. To provide distances for these sources a clustering analysis with a friends-of-friends algorithm was run. Assuming that the sources in the ATLASGAL CSC represent the individual parts of GMCs with the highest column density, clustering them in PPV space allows for the identification of the large-scale features that contain them. Clustering thus provides a statistical way to check the individual distance assignments of all sources associated with the same structure. Within a cluster, a distance can also be assigned to sources for which the distance ambiguity could not be resolved through the HI emission analysis. The major molecular gas complexes in the Milky Way, such as W31, W43, and G305, have been studied extensively and their distances have been determined accurately. These distances are adopted for sources identified within clusters corresponding to these structures. In addition, since a large number of GMCs present strong velocity gradients, the velocities of the constituent clumps may vary greatly. These differences in velocity may result in incompatibility with the kinematic distance assignments[9] that affect their estimated physical properties and increase the scatter in their Galactic spatial distribution. Clustering is an effective way to mitigate these issues.

Clustering analysis over the ATLASGAL CSC identified 776 clusters with many corresponding to well-known star-forming regions in the Galaxy (Urquhart et al., 2018).

The full ATLASGAL catalogue with distances in physical properties as described in Urquhart et al. (2018) is available for download[10].

[9] Two clumps with similar velocities ($\Delta v < 0.5$ km s^{-1}) may be applied kinematic distances that differ by 0.5 kpc.

[10] `http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/MNRAS/`

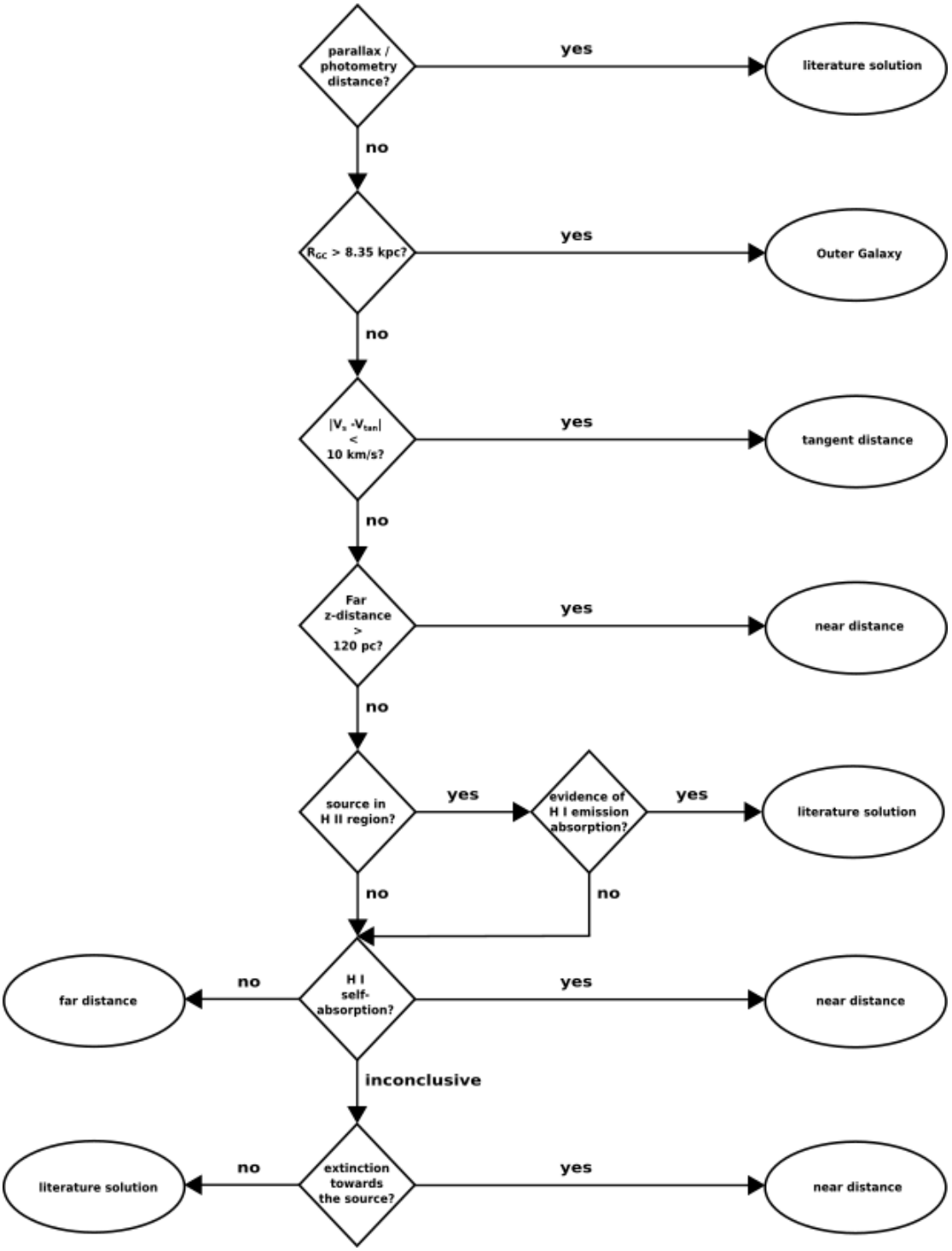


FIGURE 2.3: Flow chart of the algorithm used in Urquhart et al. (2018) to assign distances to ATLASGAL CSC clumps. Figure reproduced after Urquhart et al. (2018)

2.4 Hi-GAL

The Herschel infrared Galactic Plane Survey (Hi-GAL) is an Open Time Key Project of the Herschel Space Observatory (Molinari et al., 2010b,a). It comprises a suite of five inner Galaxy plane surveys observing from the near-infrared to radio at 70, 160, 250, 350, and 500 μm. With diffraction-limited spatial resolution and including the peak of spectral energy distribution (SED) of the cold ISM at 8K $< T <$ 50K, Hi-GAL represents an optimal tool to study luminosity, temperatures, and masses of cold gas structures from the CMZ to the Outer Galaxy. As the Herschel telescope can image multi-wavelength extended emission on scales ranging from the diffuse ISM and dense filamentary structures to compact and point-like sources (Molinari et al., 2010b; André et al., 2010), Hi-GAL data are ideal to trace the stages of the star formation process, from clouds and filaments to the collapse of dense cores into protoclusters (Zavagno et al., 2010; Elia et al., 2014; Fuller et al., 2015; Elia et al., 2017).

The Hi-GAL catalogue considered in this study refers to the first Hi-GAL data release which spans longitudes $-71° \leq l \leq 68°$ and latitudes $|b| \leq 1°$. The region of sky covered by Hi-GAL is estimated to include most of the potential star formation sites in the Inner Galaxy (with $\sim 80\%$ of YSOs being located at latitudes $|b| \leq 0°.5$). Hi-GAL is the largest Herschel observing programme to date (900 hours).

Observations in the five Hi-GAL photometric bands were acquired simultaneously over $\sim 2.2° \times 2.2°$ tiles. Each tile was observed with the Photodetector Array and Camera Spectrometer (PACS, Poglitsch et al., 2010) and the Spectral and Photometric Imaging Receiver (SPIRE, Griffin et al., 2010) in parallel mode (pMode). To mitigate the thermal drifts affecting the differential bolometers in the PACS and SPIRE arrays, the tiles were scanned twice in perpendicular directions over the two passes.

2.4.1 Data

The raw Hi-GAL data collected during the PACS and SPIRE timelines were reduced through a two-stage pipeline: the construction of the reduced maps through the ROMAGAL map-making algorithm (Traficante et al., 2011) and post-processing with the WGLS package (Piazzo et al., 2012) to rid the maps of artefacts.

A catalogue of compact cold objects extracted from the Hi-GAL data (Elia et al., 2017) is constructed by merging the Hi-GAL single-band photometry of the sources into a five-band catalogue, which is then filtered by applying constraints on sources' SEDs.

The CUrvature Thresholding EXctraction algorithm (CUTEX, Molinari et al., 2011) was used for the emission extraction in the reduced Herschel cubes. CUTEX constructs a "curvature" image of the emission by taking the second (Lagrangian) derivative in four different directions (x, y, and two diagonals). In this new image, the (slowly varying) curvature corresponding to fore- or background emission on large and intermediate scales is damped, while the curvature of point-like and compact resolved sources is amplified. The areas exceeding a curvature threshold are then considered as candidate sources. The CUTEX algorithm implements a two-dimensional Gaussian profile fit to estimate the sources' integrated flux. The extraction is performed in all five Herschel bands. A multi-band catalogue is then constructed by associating to each source its image in all bands. The matches are produced by iteratively checking the positions of the sources in two adjacent Herschel bands (Elia et al., 2010, 2013). An assignment is made when the centroid of the source at the shorter wavelength is contained in the ellipse approximating the source at the longer wavelength. When a source has more counterparts at the short wavelength, a unique association is established by selecting the short-wavelength counterpart with the shortest distance to the long-wavelength ellipse centroid. The unassigned short-wavelength sources are labelled as 'independent catalogue entries' and are checked for counterpart matching at shorter wavelengths. This merging algorithm yields a catalogue in which each entry corresponds to a source with up to five detections associated with it.

The full catalogue construction, its caveats, and the determination of the other physical properties appearing in it are found in Elia et al. (2017). The Hi-GAL physical catalogue for the Inner Galaxy is available for download from the VIALACTEA Knowledge Base[11].

[11] `http://vialactea.iaps.inaf.it/vialactea/public/HiGAL_clump_catalogue_v1.tar.gz`

Chapter 3

Cloud extraction: Data and methods

This chapter describes the methods that were employed to construct a cloud catalogue of the CHIMPS survey based on a SCIMES emission segmentation. A brief description of the SCIMES and FellWalker paradigms is given at the beginning. The former is a watershed algorithm widely used for segmenting multidimensional emission data arrays, while the latter is a more recent method that provides image segmentation based on dendrograms and clustering theory (these methods are described in detail in Appendices A and B). SCIMES relies on the natural transitions in the emission to produce a physics-oriented catalogue of emission structures (Colombo et al., 2015a). This approach represents an evolution over pixel-based cloud segmentation methods for which any property of the ISM can be chosen for data segmentation. Thus a comparison between FW and SCIMES represents a step forward in understanding method-dependent biases in survey results. A brief introduction to these algorithms is followed by the description of the preparation of signal-to-noise datacubes on which the emission extraction is performed. Finally, a post-processing routine is introduced to uniquely identify clouds in the overlapping areas between adjacent CHIMPS regions (see section 2.1 and Table 4.1). To implement a comparison with a different tracer, a subsample of COHRS at the intersection with CHIMPS is selected. The COHRS catalogue is constructed with a SCIMES segmentation with different values of the dendrogram defining parameters. Crucial to the construction of any catalogue of the physical properties of GMCs in the

determination of their distances. For this purpose, a combination of different methods and surveys is used.

3.1 The FellWalker algorithm

The FellWalker (FW) algorithm implements a variation of the watershed paradigm. While watershed algorithms perform segmentation by recognizing regions of low emission around local minima (catchment basins) and tracing out the boundaries (watershed lines) that separate them (Roerdink & Meijster, 2001), FW first searches for local maxima. Each partition of the dataset is identified through gradient tracing and associated with its corresponding maximum. This procedure resembles the HOP algorithm, a method devised to find groups of particles in N-body simulations (Eisenstein & Hut, 1998). HOP and FW share a similar design, which makes them more sensitive to the variation of the baseline threshold than to their other parameters. The FW design aims to overcome the issues arising in algorithms based on the analysis of contour levels (CLUMPFIND being a clear example, see Williams et al., 1994, and Appendix A). The FellWalker strategy determines the paths of the steepest ascent originating at each data point with an emission value that exceeds a given baseline threshold. The set of voxels belonging to all paths associated with the same peak is then identified with an individual cloud in the emission data array. The emission extraction is regulated by a set of configuration parameters that define the data value below which pixels are considered to be in the noise, the minimum dip between two adjacent peaks for them to be considered separate emission features, and the minimum number of voxels in a peak to be considered an independent source (see Appendix A for the full list).

3.2 SCIMES

The Spectral Clustering for Interstellar Molecular Emission Segmentation (SCIMES) is a segmentation algorithm that implements a clustering process based on graph theory. Clustering is an unsupervised technique used to classify patterns by dividing a set of data into groups (clusters). Data points that belong to the same cluster are more similar to one another (with respect to some of their properties) than to the points grouped

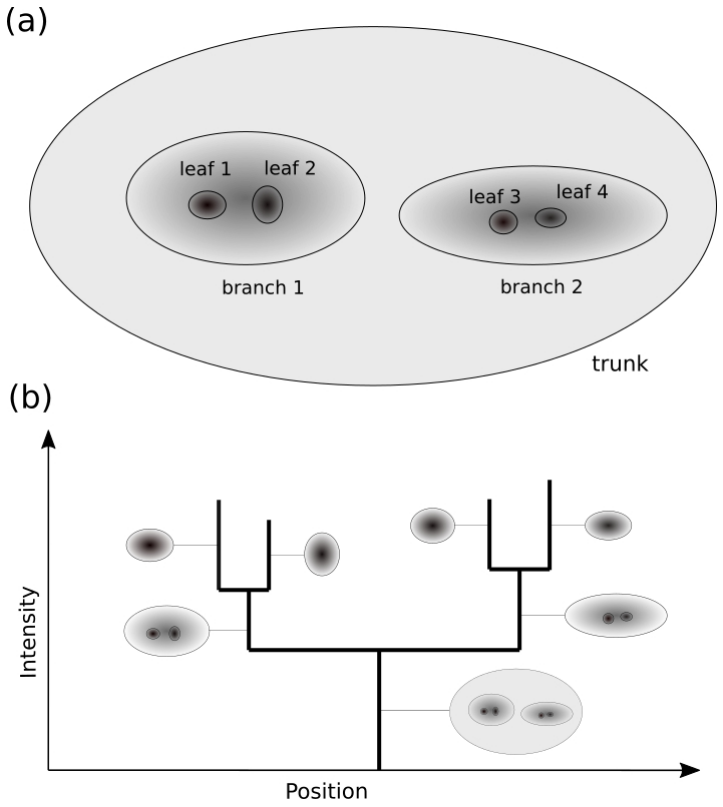


FIGURE 3.1: Example of molecular cloud emission (a) and associated dendrogram (b). Darker colour indicated higher intensity of emission.

into different clusters (Jain et al., 1999). In the framework of a clustering problem,

finding molecular clouds in a PPV datacube or an image is translated to the process of clustering pixels that are considered as part of individual entities.

The global hierarchical structure within a molecular line datacube is encoded into a dendrogram. Each point of the dendrogram can be intuitively identified as defining an isosurface at a fixed emission level. In this framework, leaves represent three-dimensional contours (or isosurfaces at given emission levels) that contain a single local maximum (see Figure 3.1). Leaves are the top level of the dendrogram. The branches of the dendrogram are vertical and horizontal lines that join two or more leaves with length (of the vertical segments) proportional to the range of contour levels across which the properties of the emission do not change significantly with respect to some chosen similarity criterion (Rosolowsky et al., 2008).

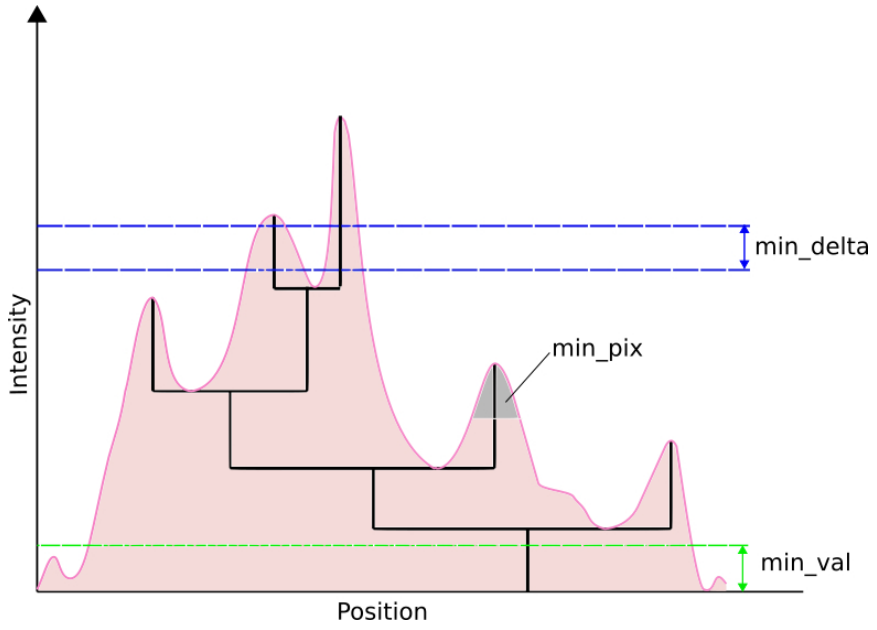


FIGURE 3.2: Schematic construction of a dendrogram for a one-dimensional emission profile. The dendrogram diagram reproduces the emission structure as a function of the contour level.

The emission dendrogram is constructed by identifying the voxels with the largest emission value within a box of a given size. Then, the elimination of local maxima proceeds as shown in Figure 3.2. Peaks are removed if their emission is below the set `min_val` or when they belong local to an isosurface with a volume smaller than a specified number of voxels (`min_pix`). If the difference between the peak and the value of the emission at the contour level where it merges with a neighbouring peak is smaller than a threshold value (`min_delta`) both contour profiles are counted as a single local maximum. The contour level at which two isosurfaces merge is called a merger level. At lower emission levels, all the branches and leaves eventually merge into the trunk of the tree structure.

Dendrograms can also be seen as mathematical graphs by considering the leaves as the vertices of the graph. The edges of the graph can be weighted using the properties of the highest-level isosurface containing each pair of leaves. Those weights are collected into similarity matrices and passed to the spectral clustering algorithm. Spectral clustering employs the eigenvectors of the Laplacian matrix to perform a dimension reduction and construct a metric space in which data points with similar emission properties are collected in separate regions. These sets of points are the independent objects identified as the ("molecular gas") clusters in the PPV data set.

Spectral clustering produces optimal cuts of the structure tree, which identifies the molecular clouds while respecting the hierarchy of the dendrogram structures.

The SCIMES method has proven to be robust under changes in the dendrogram-defining parameters and different noise realisations Colombo et al. (2015a). The SCIMES segmentation results are stable when the spatial resolution is degraded up to a factor of 10. Coarse resolutions (> 10 pc) affect the algorithm performance (Colombo et al., 2015a). Thus, SCIMES performs best in complex environments, making it an optimal choice for cloud identification in high-resolution Galactic plane surveys.

SCIMES expands the friends-of-friends paradigm by introducing neighbourhoods defined by the physical properties of the emission structure. The volume criterion (see Appendix B) was found to produce a better clustering performance. In theory, similarity relations (matrices) can be defined using any property of the ISM, including star formation rate and metallicity. This new definition of neighbourhood also broadens the very concept of molecular cloud to the more general 'molecular gas cluster. Colombo et al. (2015a) define molecular gas clusters as a category of discrete objects within the molecular ISM that have common physical properties and can be segmented by a well-defined set of similarity criteria. This category includes molecular clouds.

3.2.1 An example - The Orion-Monoceros region

Figures 3.3 and 3.4 depict an example of the SCIMES segmentation of the ^{12}CO (1-0) emission in the Orion-Monoceros region. The data set[1] used for the segmentation was obtained with the 1.2-m millimeter wave telescope at the Harvard–Smithsonian Center for Astrophysics (Wilson et al., 2005). The set has a spatial resolution of 8.4 arcmin which, at the average distance of the complex ($\sim$ 450 pc), corresponds to $\sim$ 1 pc. The images below (Figure 3.3) span $200 \times 160\mathrm{pc}^2$, while the data cube has a velocity resolution of 0.65 $\mathrm{km\,s^{-1}}$ with velocities ranging from -3 to and 19.5 $\mathrm{km\,s^{-1}}$. The data have a sensitivity σ_{rms} of 0.26 K.

Both Figures were constructed following the SCIMES tutorial[2].

[1] The dataset was obtained from `https://www.cfa.harvard.edu/rtdc/CO/NumberedRegions/DHT27/index.html`.

[2] `https://scimes.readthedocs.io/en/latest/tutorial.html`

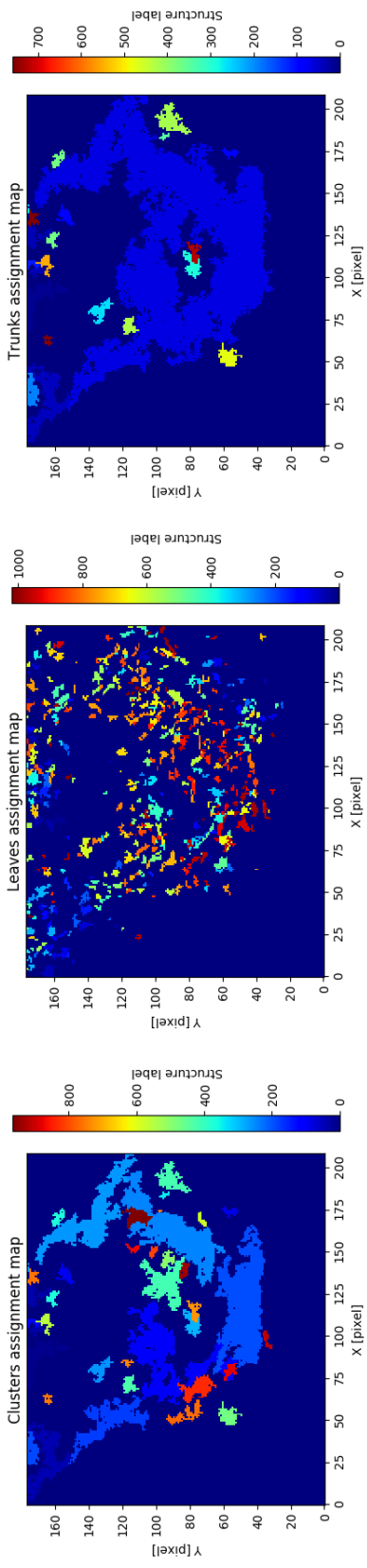


FIGURE 3.3: The projected clusters (left), leaves (centre), and trunks (right) assignment cube that were constructed during the SCIMES segmentation of the Orion-Monoceros complex. The colours correspond to the label assigned by the algorithm.

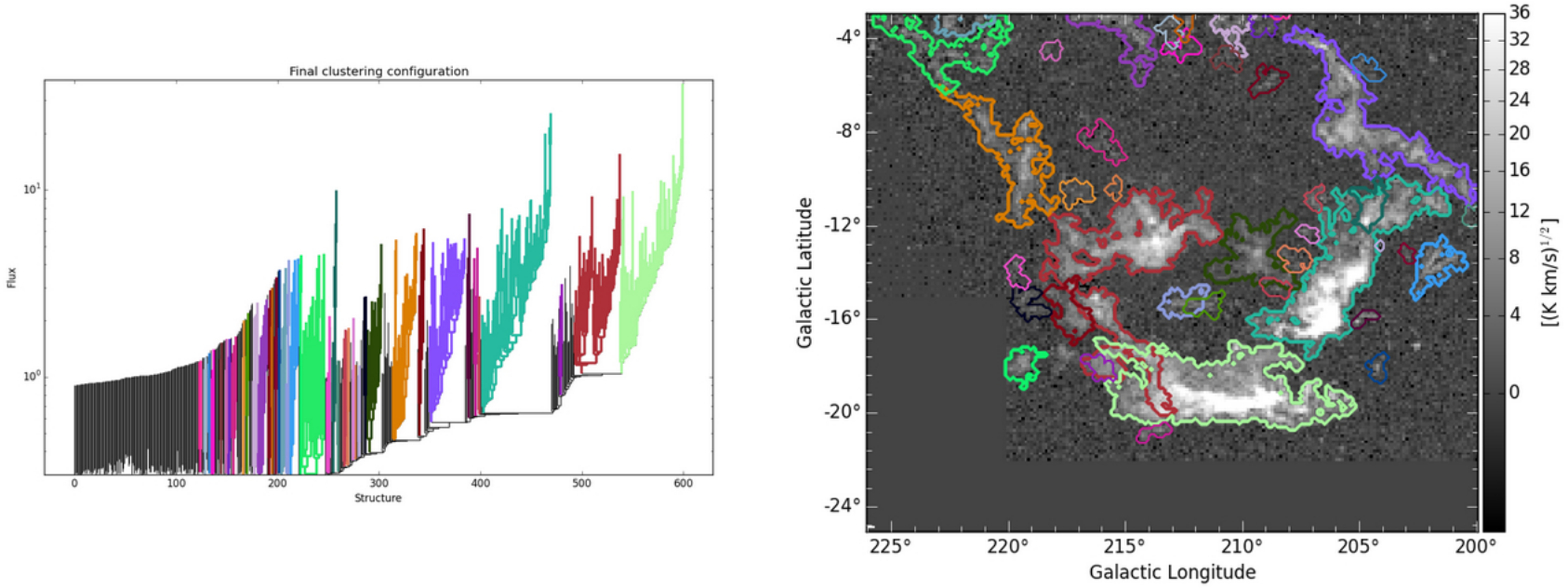


FIGURE 3.4: The dendrogram (left panel) and its corresponding projected segmentation (right panel) of the Orion-Monoceros region. The coloured contours in the right panel refer to the corresponding clusters identified as "cuts" (partitions) in the emission dendrogram. Each cluster spans several levels of the emission substructure, covering the range of emission levels (isosurfaces) that characterise its branches and leaves.

3.3 Data preparation

In this analysis of the difference between the FW and SCIMES extraction algorithm, we consider the ($J = 3 \rightarrow 2$) emission from the reduced data in the 10 regions constituting the CHIMPS survey (see section 2.1.1 and Figure 3.5). Before running the SCIMES extraction, CHIMPS data are prepared following the recipe used by Rigby et al. (2019) for the FW extraction. The reduced data are spatially smoothed to a resolution of 27.4 arcsec (resulting from the application of a 3-pixel FWHM Gaussian filter) to increase the signal-to-noise ratio (SNR). The smoothed data have rms values of $0.09^{+0.03}_{-0.03}$ K per 0.5 km s^{-1} channel. This value is the median of the distribution with uncertainties corresponding to the first and third quartiles (Rigby et al., 2019). Because of the variable weather conditions and the varying number of active receptors during the four years of observations, the original CHIMPS datacubes do not present a completely uniform sensitivity across the entire survey (Rigby et al., 2016). To avoid loss of good signal-to-noise sources in regions of low background and to prevent high-noise regions from being incorrectly identified as clouds, the source extraction is performed on the SNR cubes instead of brightness-temperature cubes. An SNR map is created from an existing brightness temperature cube by dividing it by the square root of its variance component. The resulting data array measures the SNR at each voxel of the original cube [3]. This operation is performed by the MAKESNR package of KAPPA in the Starlink suite. This approach was applied to continuum data in the JCMT Plane Survey (JPS) by Moore et al. (2015) and Eden et al. (2017), who noted that this method produced the best extraction results. Finally, the background noise is identified and subtracted from the SNR cubes by applying the Findback filter with a set neighbourhood with a side of 50 voxels (see Appendix E).

[3] http://starlink.eao.hawaii.edu/docs/sun95.htx/sun95ss108.html#Q1-135-550

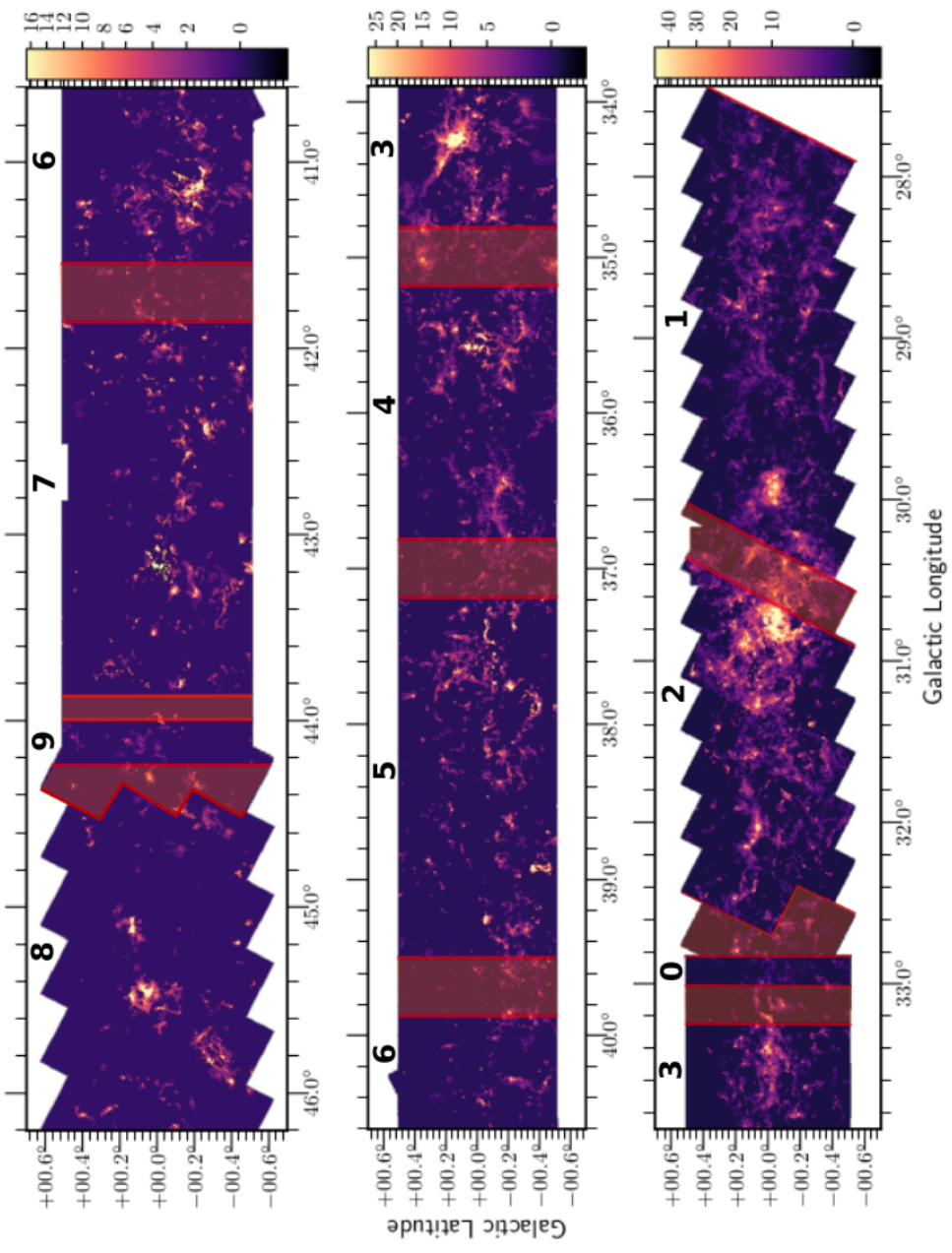


FIGURE 3.5: Integrated intensity map ($\int T_A^* dv$) of CHIMPS (full survey). The colour bar shows the scaling in units of K $\mathrm{km\,s^{-1}}$. The 10 regions into which the survey is divided are delimited by red lines. Orange shading denotes the overlapping areas between adjacent regions. Region numbers are printed above the map.

3.4 Emission extraction

The SCIMES parameters are defined as multiple of the background σ_{rms}. For signal-to-noise cubes, $\sigma_{\mathrm{rms}} = 1$ by definition. For each region, the SCIMES parameters are set to generate an emission dendrogram in which emission below $5\sigma_{\mathrm{rms}}$ (`min_val` $= 5\sigma_{\mathrm{rms}}$) is not considered. This minimum SNR value for a feature to be detected as a source was chosen to mitigate the occurrence of false positives (artefacts arising at low noise levels). Each branch of the dendrogram is defined by an intensity change of $5\sigma_{\mathrm{rms}}$ (`min_delta` $= 5\sigma_{\mathrm{rms}}$). This value is chosen to match `min_val` so that two adjacent peaks are considered distinct only if the difference in their values is also greater than 5. In addition, the minimum number of voxels an emission peak must contain to be included in the dendrogram (`min_npix`) is set to 16, which is at least three resolution elements worth of voxels (= 16). This value corresponds to the volume of a cubic source with a width of 2.5 voxels in each of the three axes. Lowering this threshold increases the likelihood of identifying spurious noise artefacts as features of the emission. These specific values were chosen to match the corresponding FellWalker configuration parameters (`MinHeight`, `Noise`, `MinPix`, see section A.2) used by Rigby et al. (2016) for their CHIMPS extraction.

Since the distances to the dendrogram structures are not known, the volume and luminosity affinity matrices required for spectral clustering cannot be generated from spatial volumes and intrinsic luminosities. Instead, PPV volumes and integrated intensity values are used (see Appendix C).

3.5 Post-segmentation processing

To clean the catalogues of spurious sources and noise artefacts that are left after extraction, an additional filter is applied. This mask leaves those clouds that either extend for more than 9 voxels in one direction (spatial or spectral) or that contain at least one $3 \times 3 \times 3$ voxel cube. While the former requirement ensures that also filamentary structures are considered, the latter ensures that each cloud is fully resolved in each direction (the width of the beam being 3 voxels). In addition, smaller clouds in contact with edges of the regions and those with no known column densities are removed from the catalogue.

The remaining clouds that touch the edges are flagged with an 'EDGE' label in the catalogue. Eventually, to construct the final catalogue and its corresponding assignment mask, a selection method is used to handle the clouds in overlapping areas between adjacent regions. This procedure is described below.

3.6 Overlapping areas

To avoid double-counting clouds and to account for the discrepancies in the extraction maps near longitudinal edges due to the separate dendrograms representing the gas structure in each region, the following prescription is utilised to treat objects extracted in the overlapping areas. This novel algorithm is based on the post-segmentation processing in the SEDIGISM (Duarte-Cabral et al., 2021) and COHRS (Colombo et al., 2019) catalogues. In each region, clouds within the overlapping area that crosse the longitudinal edges (clouds 3 and 4 in Fig. 3.6) are removed. Such clouds do not have closed isocontours in the region in question (Colombo et al., 2015a). These objects are recovered from the SCIMES extraction in the adjacent regions that contain the clouds to their full extent. Some regions present clouds that span the entire overlapping field. In order not to discard a significant amount of gas mass, these clouds are split at the edge of one region, assigning the portion in the overlapping area to the region that contains most of the cloud (cloud 1 in Fig. 3.6). The remaining portion of the cloud is then added to the final catalogue.

Finally, all objects that do not overlap between the regions (cloud 5 in Fig. 3.6) are included, and whenever two (or more) clouds overlap, the smaller object between the two regions is discarded. Through this procedure, a catalogue of 2944 molecular clouds is constructed. Distances to the catalogue sources are still to be determined at this stage.

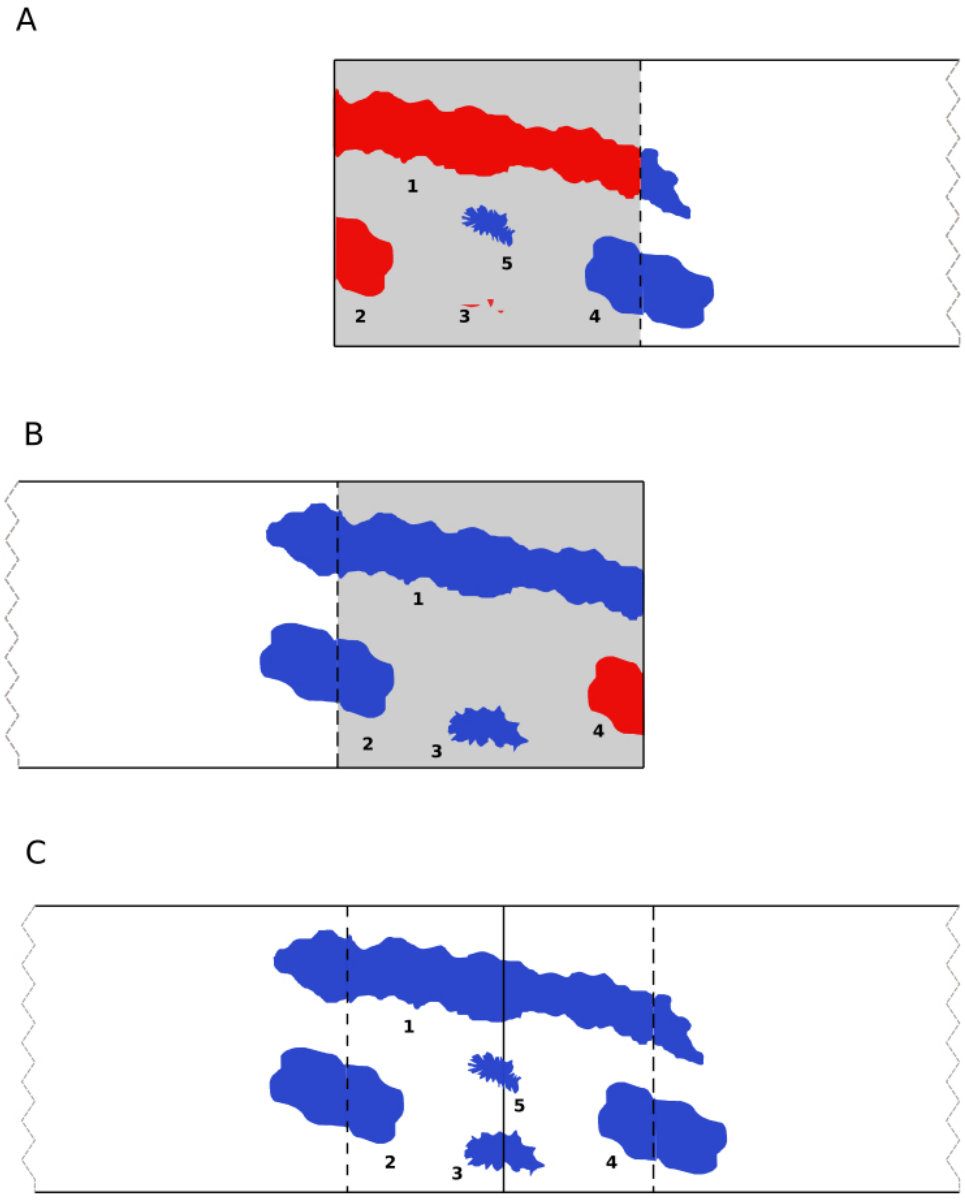


FIGURE 3.6: Prescription for cloud removal in the overlapping areas (shaded area in the panels) of adjacent regions (panel A and B). In each region, the clouds within the overlapping areas that cross longitudinal edges are removed. The clouds that are removed in each region (clouds 1, 2, and 4, drawn in red) are recovered from the other region. Clouds that span the entire overlapping area (cloud 1) are split at the longitudinal edge that marks the end of the region (panel A). The portion of the cloud contained in the shaded area is then assigned to the region that contains most of the cloud (panel B) and removed from the other (panel A). The portion of the cloud left in panel A (blue tip) is then added to the final catalogue (panel C). Whenever two (or more) clouds overlap (cloud 3), the smaller object between the two regions is discarded. All objects that do not overlap between the regions (cloud 5) are retained.

3.7 COHRS data

To implement a comparison with a different tracer, the distributions of the distances, masses, virial parameters, and densities associated to and the ^{12}CO (J = 3 - 2) emission from the COHRS catalogue are also considered. The $J = 3$ level of ^{12}CO has a temperature of 33 K, this features along with a critical density of $\sim 5 \times 10^4$ cm^{-3}, make this transition an ideal indicator of star formation of the warm material (10–50 K) of medium density (10^4cm^{-3} at 20 K) around cores heated by active star formation.

Mapping the emission of the ^{12}CO isotopologue, the extraction of molecular clouds in COHRS requires a different SCIMES parametrization (`min_delta` = $5\sigma_{\rm rms}$, `min_val` = 0K, `min_npix` = Ω_b, where Ω_b is the solid angle of the beam expressed in pixels, see 2.2 and Colombo et al. (2019)). In addition, caution should be exercised when comparing physical quantities with definitions dependent on the catalogue. The same physical quantity may be defined up to different scaling factors in different catalogues (see the virial parameter in Colombo et al. (2019) and reference for instance).

For this work, the COHRS catalogue has been reduced to those sources that fall within the area covered by CHIMPS. This reduction consists of a sub-catalogue of the COHRS fiducial catalogue, the set of COHRS sources with broadcast inaccuracy smaller than 5 voxels. Distance assignments in COHRS make use of the position of BGPS sources (Colombo et al., 2019). This sub catalogue comprises 250 sources. When the position of a BGPS object with a unique distance belongs to a SCIMES dendrogram structure, that structure/cluster inherits the BGPS distance. This assignment is defined as 'exact'[4]. Objects for which an exact assignment is not found are given a broadcast assignment. As these objects may be the substructures of larger connected emission features with exact distances, they inherit the closest distance within the larger structure. The broadcast inaccuracy measures the closest distance in voxels from the distance-assignment-position to the outer surface of the cloud. By definition, exact distance clouds have zero broadcast inaccuracy.

[4] A small fraction ($\sim$ 0.2% of the entire catalogue) of objects in COHRS possess substructures with different distance assignments. The distances of these objects are chosen to be the near distance of the brightest spot within the object. This assignment is justified by the assumption that the largest amount of cloud mass resides in its substructure.

3.8 Distances

The complex spatial distribution of molecular emission in the plane of the Galaxy makes it difficult to establish accurate distances to molecular clouds and clumps based on light-of-sight information alone. The most accurate (and model-free) distances of star-forming complexes to date were determined by parallax. An important example of this technique is provided by the distance measurements to masers obtained via Very Long Baseline Interferometry Reid et al. (2014). However, the existing distance catalogues produced by these measurements are still not exhaustive and include too few of the objects belonging to the regions surveyed by CHIMPS. When a model that mimics the Galactic rotation curve has been established, line-of-sight-velocity information provides a robust method for the calculation of kinematic distances (Brand & Blitz, 1993; Reid et al., 2014) under the assumption that the observed objects follow circular orbits around the Galactic centre. A distance assignment to the extracted SCIMES sources was constructed by combining two different catalogues and using the Bayesian distance calculator of Reid et al. (2016). First, the CHIMPS catalogue of ^{13}CO (3-2) emission extracted through the FW algorithm (Rigby et al., 2019) is considered. The main catalogue consists of 4999 sources, of which 3664 are considered robust (Rigby et al., 2019). The Bayesian distance calculator was used to estimate the possible near and far kinematic distance - and associated uncertainties - for each of the clumps (Rigby et al., 2019). No assumption about the sources being associated with spiral arms was made, and the standard Galactic rotation model (Reid et al., 2014), with a distance to the Galactic centre of $R_0 = 8.34 \pm 0.16$ kpc was adopted for the calculations. Rigby et al. (2019) then use several methods (based on geometric arguments and volumetric considerations) to discriminate between the near and far kinematic distances and make the proper assignment. This catalogue is referred to as the FellWalker (FW) catalogue. A sub-catalogue of the FW catalogue is defined by only considering the robust sources. This label indicates sources that are not false positives or single coherent sources at low S/N which are hard to discern by eye. The reduced catalogue is also free of sources consisting of diffuse gas at low S/N that may contain multiple intensity peaks, or irregular profiles (resulting from the segmentation of clouds across tile boundaries). This sub-catalogue amounts to 3664 sources.

Distances are assigned as follows. Each SCIMES cloud is matched to a set of one or more

ATLASGAL sources (Urquhart et al., 2018). The matching process is performed by first discarding all ATLASGAL objects with velocities $|v| > 2.5v_c$ where v_c is the velocity of the centroid of the SCIMES source. An area (l,b) search then follows, allowing the closest sources (Euclidean metric) that lie within a neighbourhood of radius r arcsecs centred at the centroid of the SCIMES object to be selected. The radius r is taken by adding 38 arcsec ($\approx$ 5 pixel) to the radius of the SCIMES object (Rigby et al., 2019). Next, if this search returns multiple clouds, the distance that most sources have in common is chosen. If the distances in the set vary significantly we check if any of them belongs to an ATLASGAL cluster, and assign the cluster's distance to the SCIMES cloud. SCIMES clouds that contain one single ATLASGAL source for which the distance is not available, or in the case of clusters, ATLASGAL does not provide a cluster distance, are left unassigned. The unassigned sources are compared to the reduced FW catalogue. If a SCIMES cloud contains a single FW object (emission peak) or more FW objects with the same distance, then that distance is assigned to the cloud. If a SCIMES cloud contains multiple FW sources with different distances, the distance that corresponds to the mode of the distribution of FW distances is assigned. If this distribution has no modes, the first FW source in the list is chosen.

For the remaining unassigned clouds, associations between the unassigned SCIMES sources are made using a final volumetric search. This time an ellipsoidal volume of $0.3\,\mathrm{deg} \times 0.3\,\mathrm{deg} \times 10\,\mathrm{km\,s^{-1}}$ centred at the centroid of each remaining cloud is employed to identify the closest SCIMES centroid with an existing distance assignment. The size of this volume is in agreement with the appropriate tolerance for friend-of-friends grouping (Wienen et al., 2015) and corresponds to the median angular size and maximum linewidth of molecular clouds (Roman-Duval et al., 2009).

Finally, Reid's Bayesian calculator is employed to estimate the distances of the remaining SCIMES sources with undetermined distances with a near-far probability of 0.5).

To avoid the contamination of the results due to local sources and exclude a large number of low-luminosity clumps/clouds below the completeness limit[5], only sources

[5] As the surface brightness of the objects in a survey decreases, the ability to image and identify them in data sets also diminishes. It is thus crucial to know what fraction of sources with similar characteristics and brightnesses can be distinguished in a data set. This fraction is known as completeness. The completeness of the high emission sources equals 1: we can identify all of these sources in a data set. On the other hand, some of the less bright sources, may not be detected (being too distant or embedded in the noise, for instance). For a given class of objects, the completeness limit corresponds to the magnitude at which the completeness drops below a given threshold (commonly 90%).

with heliocentric distance > 2 kpc are included (Urquhart et al., 2018).

Galactocentric distances are calculated independently. Brand & Blitz (1993)'s rotation curve is used. The angular velocity is derived from the line-of-sight velocity, $v_{\rm LSR}$ and the Galactic coordinates l and b via the relation

$$\omega = \omega_0 + \frac{v_{\rm LSR}}{R_0 \sin(l) \cos(b)}, \tag{3.1}$$

where $\omega_0 = 220\ \mathrm{km\,s^{-1}\,kpc^{-1}}$ is the Sun's angular velocity at its Galactocentric distance $R_0 = 8.5$ kpc. The Galactocentric distance of a source is then obtained by solving

$$\frac{\omega}{\omega_0} = a_1 \left(\frac{R}{R_0} \right)^{a_2 - 1} + a_3 \frac{R_0}{R}, \tag{3.2}$$

numerically, with the constants $a_1 = 1.0077$, $a_2 = 0.0394$ and $a_3 = 0.0071$ (Brand & Blitz, 1993).

3.9 Summary

This chapter covers the methodology and data treatment used for the comparison of the FW and SCIMES segmentation algorithms on CHIMPS ^{13}CO emission. It includes

- a short overview of the Fellwalker watershed algorithm (see also Appendix A),
- an introduction to the spectral-clustering-based SCIMES algorithm (fully explained in Appendix B),
- a new post-processing algorithm to "clean" overlapping regions in segmentation maps and
- a new method to assign distances to SCIMES-extracted clouds in CHIMPS that makes use of existing catalogues (ATLASGAL and the FW catalogue) and the Bayesian distance calculator (Reid et al., 2014).

Chapter 4

A new CHIMPS segmentation

The development of a wide range of automated cloud identifying algorithms based on different paradigms (see Chapter 1) has prompted the need for a direct comparison of these methods under different conditions and for the emission of different molecules. These methods are complex and testing for biases is often problematic: only a few of them have been applied to the same data set or calibrated against a common standard. In addition, cross-correlating the physical properties of individual sources between several catalogues is often a complicated task. From this viewpoint, it is thus of interest to apply different methodologies to identify and extract GMCs from the same CO survey.

In this chapter, the Spectral Clustering for Interstellar Molecular Emission Segmentation (SCIMES) algorithm is applied to identify GMCs in the ^{13}CO data-set of the ^{13}CO/C^{18}O($J = 3 \rightarrow 2$) Heterodyne Inner Milky Way Plane Survey (CHIMPS, see section 2.1). To directly compare this segmentation to the results obtained by Rigby et al. (2019) with the FW algorithm, the dendrogram defining parameters are chosen to match the FW input configuration as described in section 3.4. To extend the comparison to the properties of a different tracer, a SCIMES segmentation of the ^{12}CO($3 - 2$) emission from the CO High Resolution Survey (COHRS) is considered (where the data are available). Finally, we present a full statistical comparison between our novel SCIMES catalogue and the one published by Rigby et al. (2019).

4.1 Emission features

Figure 4.1 shows the FW and SCIMES extractions of ^{13}CO (3–2) emission in region 3 (see text and Figure 3.5)) in the 59.72 km s^{-1} velocity plane at 27.4-arcsec resolution. In the two panels, regions of space belonging to different clouds are distinguished by different colours. The most prominent difference between the two extractions lies in the relative over-segmentation of the emission in the FW panel. This is a known feature in FW extractions in which the watershed algorithm tends to break the emission into compact clumps that are accounted for as isolated features. In addition, as Rigby et al. (2019) points out, diffuse emission around the detection threshold can be identified as sets of disconnected voxels, clustered together as individual clumps. These clouds are recognizable by their very irregular shapes and they were flagged as 'bad sources' after a visual inspection in the FW catalogue (Rigby et al., 2019). Coherent sources at low SNR and areas of emission crossing the boundaries between tiles also belong to this category. The latter sources often present very irregular segmentation due to the difference in noise levels among tiles. Such discontinuities may also create small clumps that do not originate from features in the emission map, but reflect changes in the emission in adjacent channels[1]. These inconsistencies are a consequence of performing the extraction on SNR maps. Such occurrences are however small in number and the total sample is only marginally impacted.

The final catalogue published by Rigby et al. (2019) includes 4999 sources, 1335 of them were classified as 'bad sources' thought to arise from such artefacts. On the other hand, the emission extracted by SCIMES on the same velocity plane (bottom panel in Figure 4.1) is confined to fewer individual sources, generally covering larger areas than their FW counterparts. This characteristic of the SCIMES segmentation is supported by the analysis of the geometric and physical properties of its sources (see below), thus a cloud/clump is, in general, not characterised by a single maximumn emission peak (see section 3). SCIMES clusters consist of signal from different hierarchical levels of the emission dendrogram, see the Orion-Monoceros example in section 3.2.1. The fragmentation induced by FW identifies pieces of the substructure as individual entities. In the framework of SCIMES, these clumps correspond to dendrogram branches and

[1]With the FW parametrization used for the segmentation of CHIMPS data, voxels with SNR = 2 can be included in a clump, when they are directly connected to a clump with a peak SNR > 5 (Rigby et al., 2019)

subranches. Figure 4.2 compares the FW segmentation of 57.2 km s^{-1} plane of region 3 (Table 4.1) to the structure

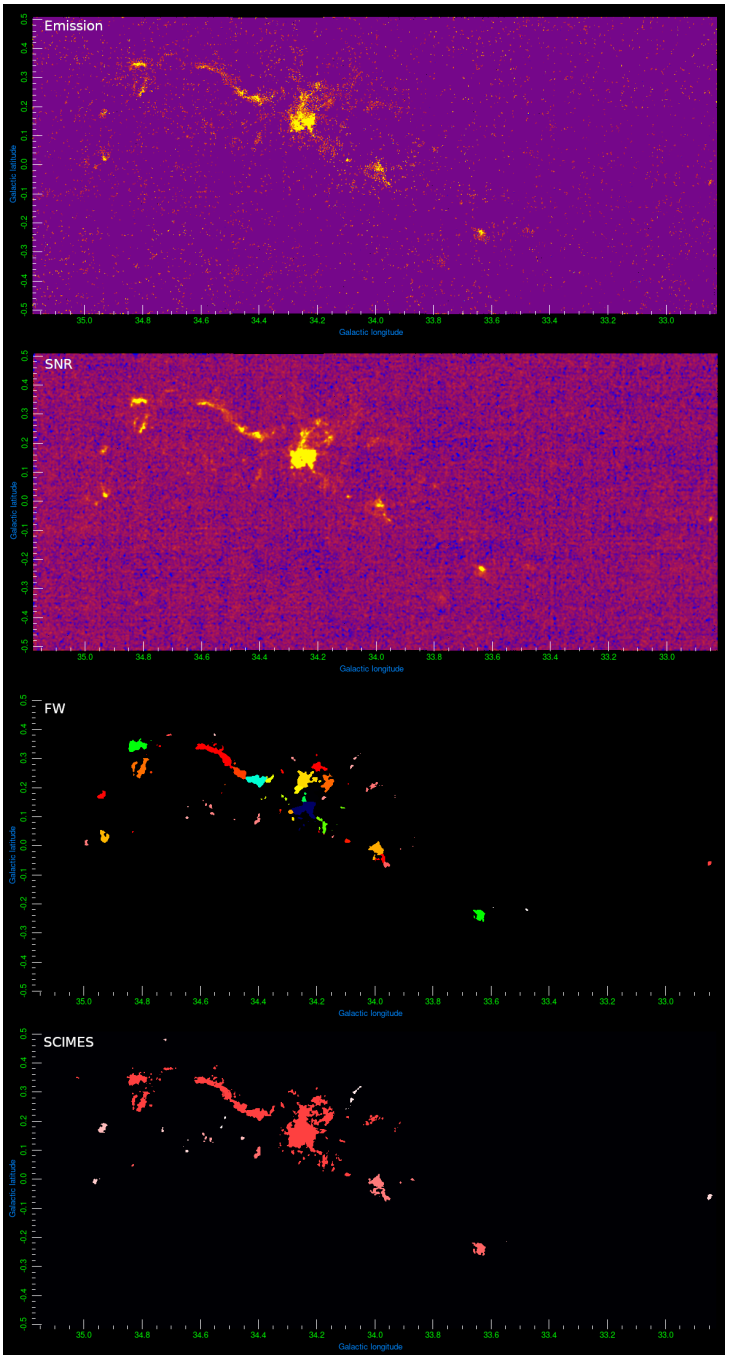


FIGURE 4.1: The top panel shows ^{13}CO (3–2) emission in region 3 (see text) in the 59.72 km s^{-1} velocity plane at 27.4-arcsec resolution. The second panel from top depicts the same velocity plane in the SNR cube. The third and fourth panels display the corresponding FW and SCIMES clusters in that plane. In both panels, different colours represent different clouds.

identified by the dendrogram leaves in the same plane. Whereas lower emission from diffuse gas causes fragmentation in the FW paradigm. The introduction of "artificial boundaries" cutting through areas of less intense emission between peaks is a consequence of the watershed algorithm. This algorithm in fact characterises disjoint clouds by single individual peaks. The volume and luminosity criteria defining SCIMES dendrograms allow for the inclusion of emission from the both hot cores and their tenuous surrounding envelopes (third panel from the top in Figure 4.1) into a single object. Furthermore, these similarity criteria (see Appendix C) allow bypassing the impact of SNR discontinuities at the edges of adjacent tiles.

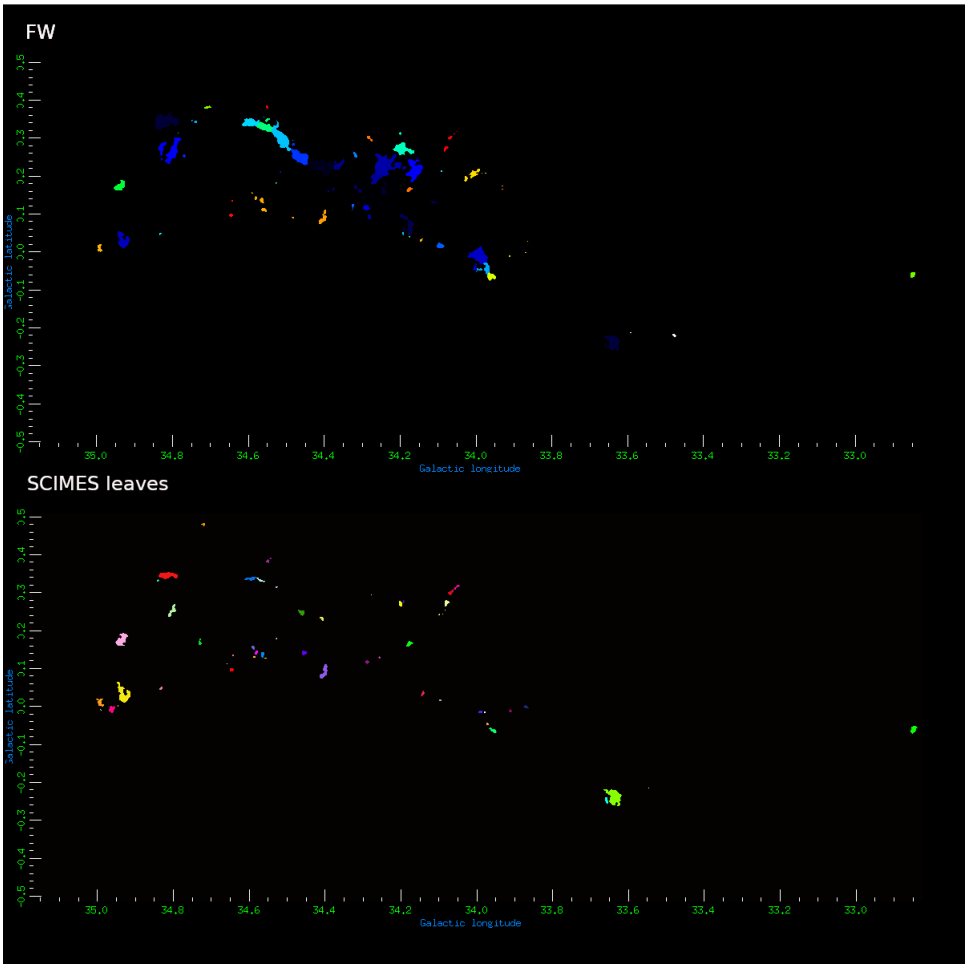


FIGURE 4.2: The top panel the FW segmentation of the ^{13}CO (3–2) emission in region 3 in the 59.72 km s^{-1} velocity plane at 27.4-arcsec resolution (see Figure 4.1). The bottom displays the dendrogram leaves in the SCIMES segmentation over the same velocity plane.

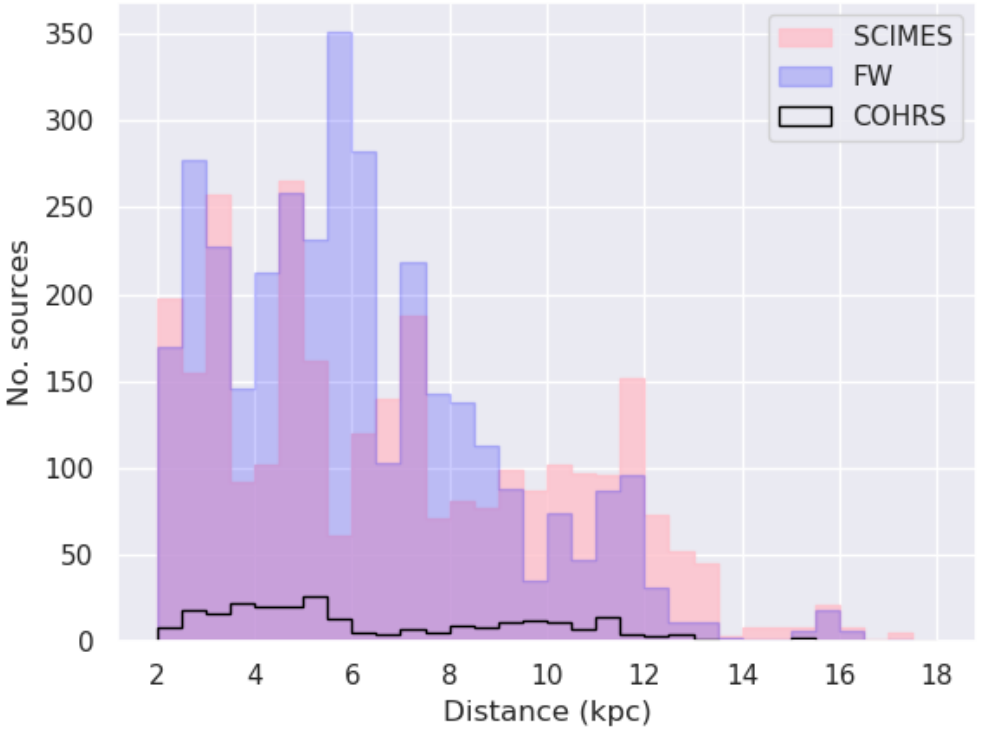


FIGURE 4.3: Distribution of heliocentric distances for the CHIMPS ^{13}CO (3 - 2) sources extracted through the FW and SCIMES segmentations. The black histogram is the distribution of sources in a subset of the COHRS catalogue (see 3.7).

4.2 Distances

Figure 4.3 shows the distribution of distances to CHIMPS ^{13}CO sources extracted with both FW and SCIMES. For comparison, the distance distribution of the subsample of COHRS sources described in section 3.7 is included. The absence of a one-to-one correspondence between FW and SCIMES clouds it impossible to establish a unique matching criterion between the FW and SCIMES distances assignment of each cloud[2]. Biuniqueness between FW and SCIMES catalogues of source cannot be established because of the different sets of structures that the two algorithms identify in the same emission feature. In assignment method described in section 3.8, a distance is assigned to a SCIMES cloud based on the FW sources it contains. The ranges of FW distance in each SCIMES clouds are plotted in Appendix F.

[2]Each FW source belongs to either one single SCIMES source or not. Each SCIMES source may contain one or more FW source or none.

To check for near-far blends in among SCIMES clouds, the catalogue was binned into two-velocity channels wide bin (1 $\mathrm{km\,s^{-1}}$). Clouds within the same bin are potentially subjected to the kinematic distance ambiguity when their distance is not well established and the assignment occurs by the Bayesian method (Reid et al., 2014). Near-far blends may thus arise for overlapping (along the line of sight) clouds within the same bin population. Only 8 of such clouds were found in the SCIMES catalogue, the projected overlapping area amount to 1010 pixels.

The different in the numbers of clouds at large distances ($\sim$ 12 kpc) and at $\sim$ 5 kpc in the FW and SCIMES catalogues are a consequence of the differences in the segmentations and of the assignment scheme of section 3.8. The distance-assignment algorithm first assigns ATLASGAL distance and then check for FW sources included in SCIMES clouds. The larger number of clouds see in the SCIMES catalogue at 12 kpc arises from those assigmnets that do not involve FW distances. The FW distance assignments are described in Rigby et al. (2019).

The top-down view of the locations of the CHIMPS sources extracted by SCIMES and FWon the Galactic plane is shown in Figures 4.4 and 4.5.

No sources closer than 3.5 kpc from the Galactic centre are found as the CHIMPS data do not probe sufficiently central longitudes (see section 2.1). The Galactocentric distribution in Figure 4.6 displays large peaks at $\sim$ 4.5 kpc and $\sim$ 6.5 kpc. These are the location of the Scutum and Sagittarius arms. The smaller peak at $\sim$ 7.5 kpc instead corresponds to the Perseus arm. Part of the Scutum arm traverses the tangential distance (see section 2.3) and the sources in this area become clustered at a distance of $\sim$ 7 kpc (Figure 4.4). The gap on the far side of the tangent points is almost absent in the distribution of Heliocentric distances (Figure 4.3) since it coincides with the region where the far Sagittarius arm begins, the far side of the Scutum arm, and with the position of the inter-arm material that accounts for the peak between 8 and 10 kpc. Similarly, the peak at 12 kpc originates from the far Sagittarius arm and the Perseus arm. The Outer arm is identified by the small peak at $\sim$ 16 kpc.

Figure 4.7 shows the heliocentric distances of the sources in the FW and SCIMES segmentation of the emission in CHIMPS and the selected sources in COHRS.

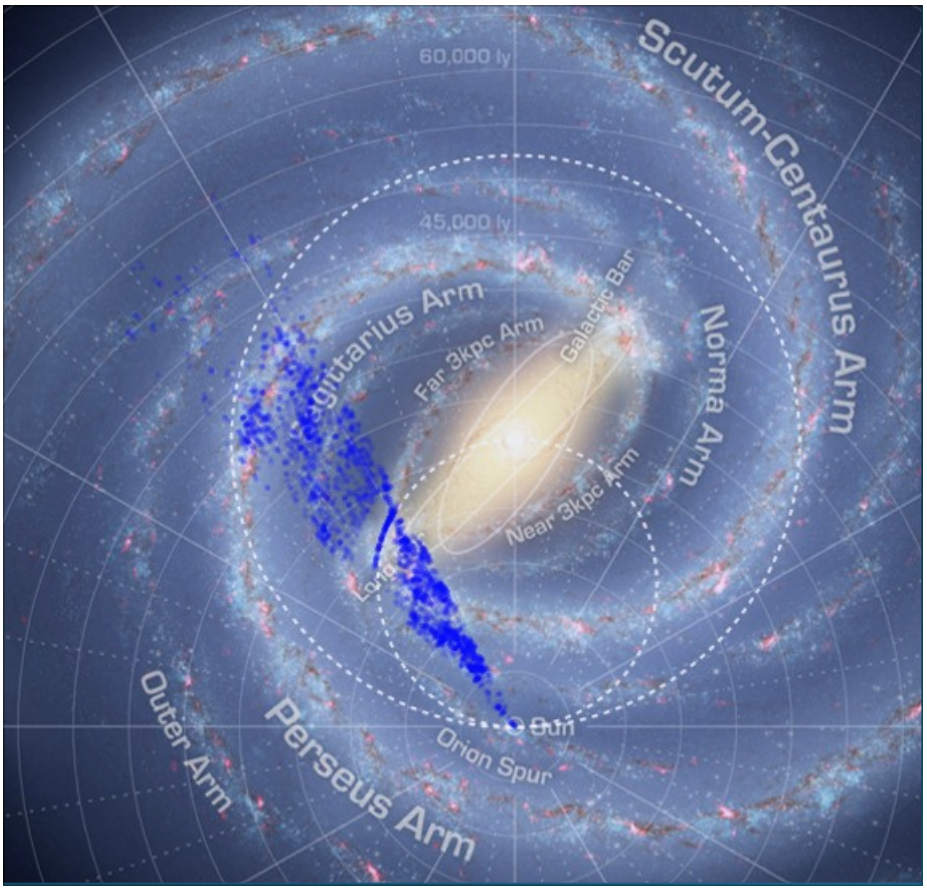


FIGURE 4.4: Top-down view of the locations of the ^{13}CO (3 - 2) extracted through the SCIMES algorithm from CHIMPS. The background image is published by (Churchwell et al., 2009b). The Solar circle and locus of the tangent points have been marked as the white dashed, and dotted lines, respectively.

To check if the FW and SCIMES distance distributions differ significantly, a Kolmogorov-Smirnov test is performed. Following the convention set in kstest in the package Scypy[3] with the null hypothesis that the two samples (distributions) are drawn from the same distribution, while the alternative is that they are independent. The test returns $k = 0.17$ with p-value $<< 0.001$, the null hypothesis can thus be rejected. The distance assignment algorithm assigns enough distances that do not depend on the FW catalogue to yield an independent distribution of distances. The same result is obtained for the distributions of independently-estimated Galactocentric distances ($k = 0.17$ with p-value $<< 0.001$).

Different features in the distance distributions also highlight the fact that there is no

[3] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html`

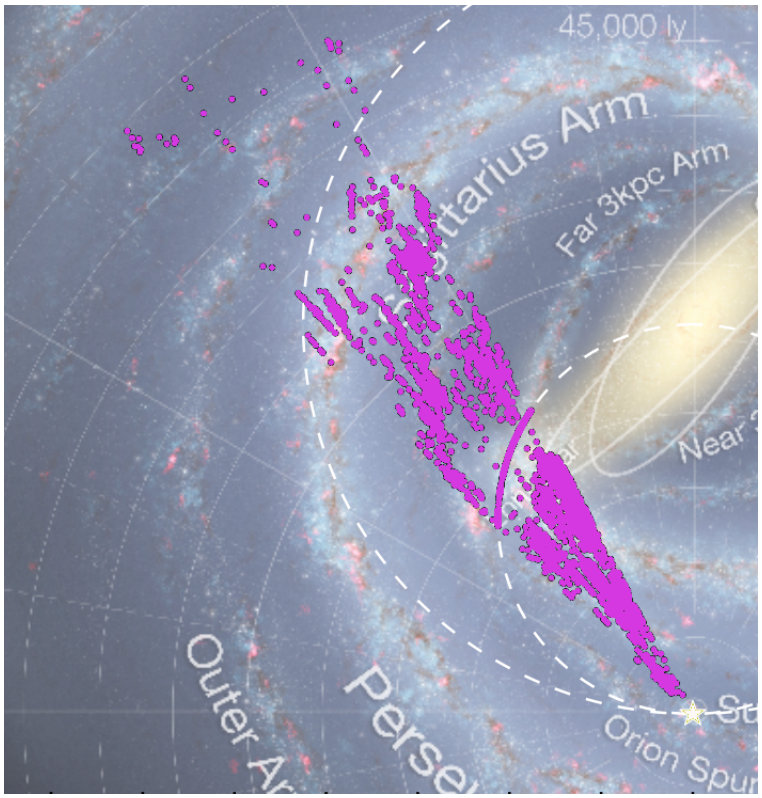


FIGURE 4.5: Top-down view of the locations of the ^{13}CO (3 - 2) extracted through the FW algorithm from CHIMPS.

one-to-one correspondence between SCIMES and FW sources. Different distance assignments (and the inexactness of the assignment) alter derived parameters and properties for individual clouds. However, these different are mitigated when the ensemble statistical properties of the sample are considered. An example is the distribution of mass derived from random distance assignments (Appendix D) resembling the distribution obtained through the distance assignment algorithm of section 4.4.1).

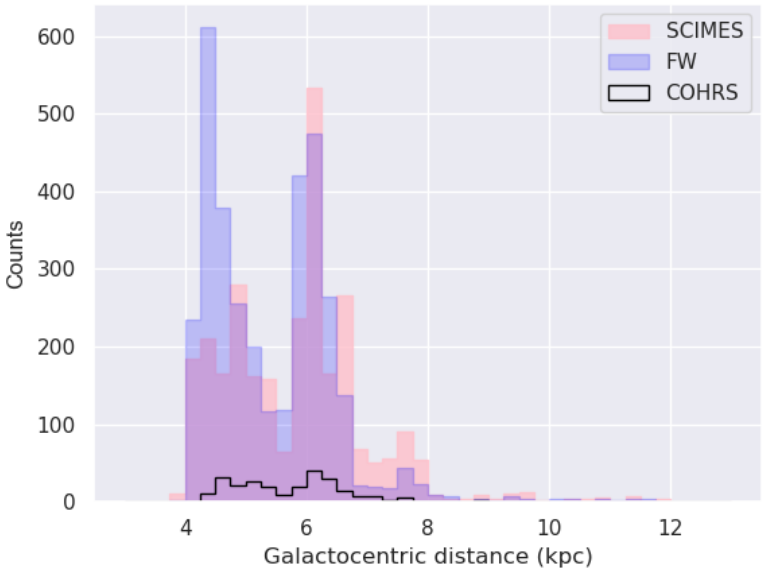


FIGURE 4.6: Distribution of Galactocentric distances for the sources extracted through the FW and SCIMES and for the sources in the COHRS subsample (see text).

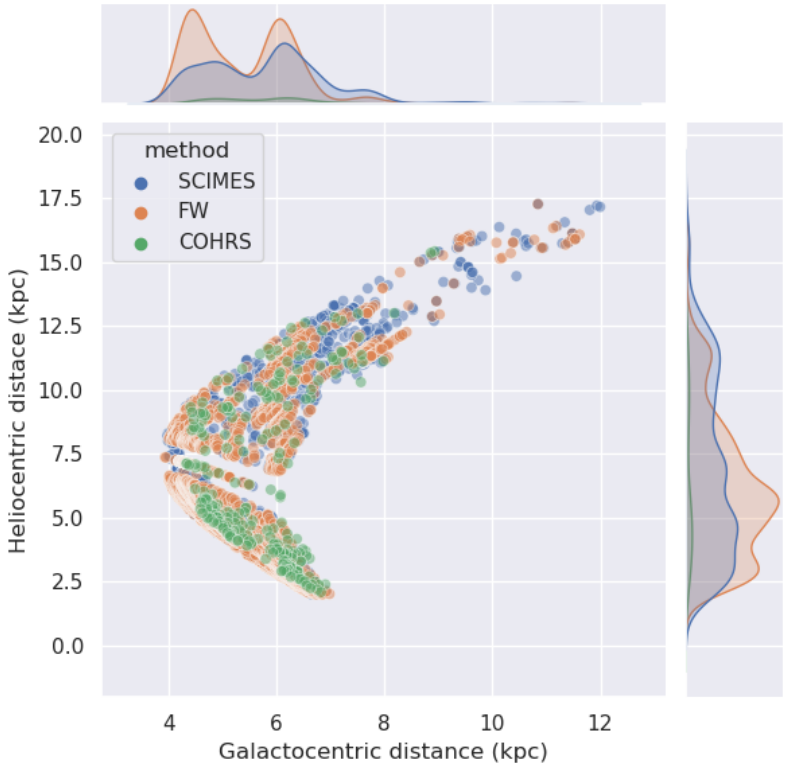


FIGURE 4.7: The heliocentric distances of the CHIMPS and COHRS sources as functions of their Galactocentric distance. The colours refer to the method of extraction and survey.

4.3 Geometry

To establish a comparison between the FW and SCIMES catalogue, the physical and geometric quantities investigated by Rigby et al. (2019) are considered. The following sections reproduce the results presented in Rigby et al. (2019) with the addition of the corresponding SCIMES quantities. Data from the COHRS fiducial sample have also been added when available.

4.3.1 Note on probability densities

In the Figures that follow, some histograms are labelled as 'probability densities'. In these histograms, each bin displays the bin's raw count divided by the total number of counts and the bin width

$$\text{density} = \frac{\text{counts}}{\text{counts} \times \text{binwidth}}. \tag{4.1}$$

The area under the histogram then integrates to 1[4]:

$$\text{area} = \text{sum}(\text{density} \times \text{binwidth}) = 1. \tag{4.2}$$

The height of the histogram bars thus also depends on the width of the bins. Notice that, by definition, bin widths smaller than unity produce bar height greater than 1.

4.3.1.1 Radii

The size of the CHIMPS clouds is defined through their 'approximate' radii. Two different radii are associated with each cloud to emphasize different characteristics of the emission. The equivalent radius R_{eq} defined as the radius of the circle whose area is equivalent to the projected area of the source,

$$R_{\text{eq}} = d\sqrt{A/\pi}, \tag{4.3}$$

[4] https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html

where d is the distances assigned to the source.

The second radius is associated with the extent of the projected cloud in the l and b directions:

$$R_\sigma = d\sqrt{\sigma_l \sigma_b}. \tag{4.4}$$

The radius R_σ is the geometric mean of the intensity-weighted rms deviations in the l and b axes (σ_l and σ_b), deconvolved by the telescope beam, and d the assigned distance. Thus R_σ depends on the emission profile of the source. In addition, R_σ is less affected by the variations in the noise level in different areas of the survey. Following Rigby et al. (2019), both radii are used in the calculation of the radius-dependent quantities associated with the SCIMES and FW extractions.

Since the dendrogram statistics implemented in SCIMES by Astrodendro do not allow for the direct computation of σ_l and σ_b, but only produces estimates of the major and minor axes of the ellipse approximating the projection of the clouds onto the coordinate plane, the calculation of R_σ is complicated by the lack of knowledge of the orientation of the ellipse with respect to the frame of reference. In this case, the conversion factor η is adopted,

$$R_{\rm eq} = \eta R_\sigma \tag{4.5}$$

to compute R_σ. The constant η is set to 2, this value corresponds to the median value found by (Rigby et al., 2019) for the FW extraction and it is a compromise between commonly-used conversion $\eta = 1.9$ (Solomon et al., 1987; Rosolowsky & Leroy, 2006; Colombo et al., 2019) and $\eta = 2.1$, the median value we found using the alternative version of R_σ

$$R_\sigma = d\sqrt{\sigma_{\rm maj}\sigma_{\rm min}}, \tag{4.6}$$

easily obtainable from the Astrodendro statistical tools for the major and minor axis of the projected SCIMES sources. For the physical properties defined below, the definitions provided in Rigby et al. (2019) were adopted. The equivalent radius is $R_{\rm eq}$ is used in all instances in which the radius enters the definition of a physical quantity. Rigby et al.

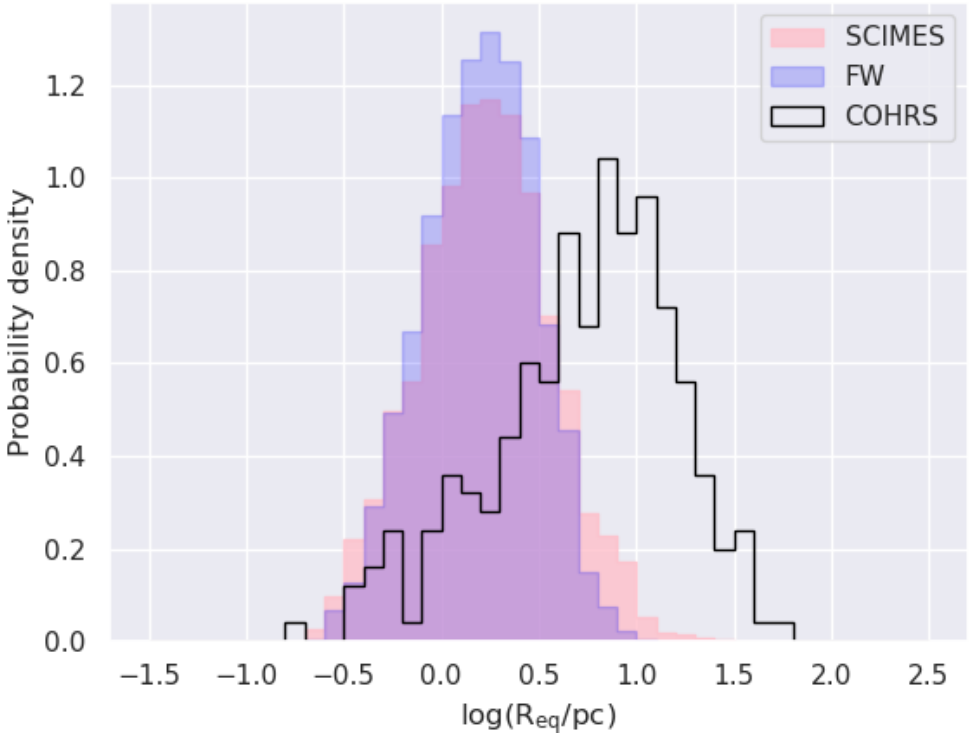


FIGURE 4.8: Distributions of $R_{\rm eq}$ of the CHIMPS ^{13}CO (3 - 2) sources in the FW (blue) and SCIMES (red) catalogues. The black histogram is the distribution of the equivalent radii of the ^{12}CO (3–2) sources in a subset of the COHRS catalogue (see text).

(2019) also used the conversion factor η (see scaling relations in Figures 4.32, 4.16, and 4.22) in definitions where comparison to different datasets required the use of R_σ. The same notation as in the original article is maintained throughout this chapter. Although the equivalent radius depends on the distances assigned to the individual clouds, the inexactness of the individual distances is mitigated when properties pertaining the entire population are considered. Appendix D provides an example of the impact of random distance assignment on the distribution of masses in the SCIMES segmentation. The right tail of the SCIMES distribution of $R_{\rm eq}$ in Figure 4.8 is an indication of the higher number of larger projected clouds extracted by SCIMES.

To check if the FW and SCIMES distribution of equivalent radii differ significantly, a Kolmogorov-Smirnov test is performed. The test yields $k = 0.067$ with p-value $= 1.04 \times 10^{-6}$ establishing that the null hypothesis of the two samples being drawn from the same distribution can be rejected.

4.3.2 Volumes

The distribution of cloud volumes (expressed as the number of voxels the cloud spans, see Figure 4.9), also reflects the fact that the SCIMES segmentation comprises both bigger and smaller clouds than its FW counterpart. Table 4.1 reports the average size of the clouds in the FW and SCIMES distributions over 10 longitudinal cuts within the CHIMPS survey (see Figure 3.5).

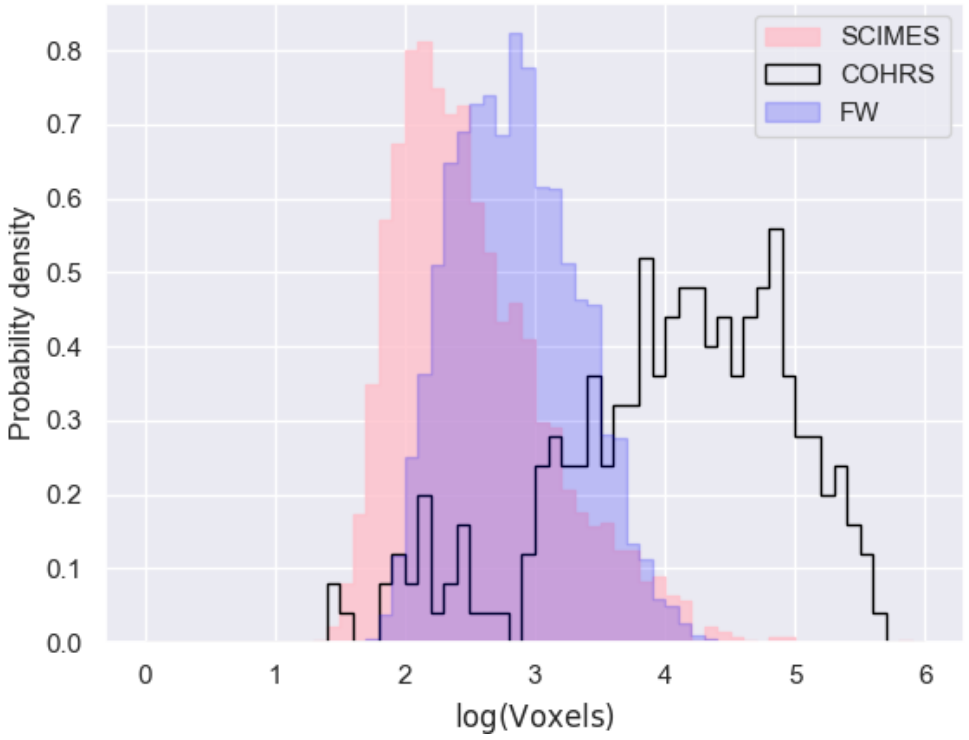


FIGURE 4.9: Distribution of volumes of the clouds in the FW, SCIMES, and COHRS segmentations.

The segmentation of the reduced COHRS fiducial sample produces clouds of larger sizes according the higher abundance of ^{12}CO and thus its ability to trace more rarefied warm gas envelopes surrounding the denser cold molecular clouds.

In general, the SCIMES segmentation comprises clouds of both bigger and smaller volumes than its FW counterpart as shown in Table 4.1. From the inspection of the regions in which SCIMES finds clouds with average volumes that are smaller than those extracted by FW (region 7 in Figure 4.10 is an example), it emerges that SCIMES finds

a small number of large sources (comprising multiple FW clumps), leaving small fragments that are not included in the large agglomerations. By number, small fragments or smaller isolated sources constitute the majority of the SCIMES clouds. These diverse emission feature are expected to arise from the application of different paradigms to regions with different spatial distribution of the emission. In the SCIMES paradigm, the shape of the emission dendrogram is determined by the spatial distribution of emission over the entire region. This is especially noticeable in the overlapping areas that adjacent regions share. The difference in shape and number of clouds in these areas requires the selection process described in section 3.5. In the smaller regions (e.g. 0 and 9) in Table 4.1, the choice of treatment of overlaps may greatly influence the average values that characterise the distribution of molecular clouds in the region. On the other hand, the FW approach is expected to be less sensitive to the overall distribution of emission as FW 'constructs' the extracted cloud locally, around emission peaks.

The comparison between the distributions of both cloud volumes and equivalent radii highlights the different results of the source extraction methods both on the complexity of the structure of molecular clouds and the environment that characterises a source's location.

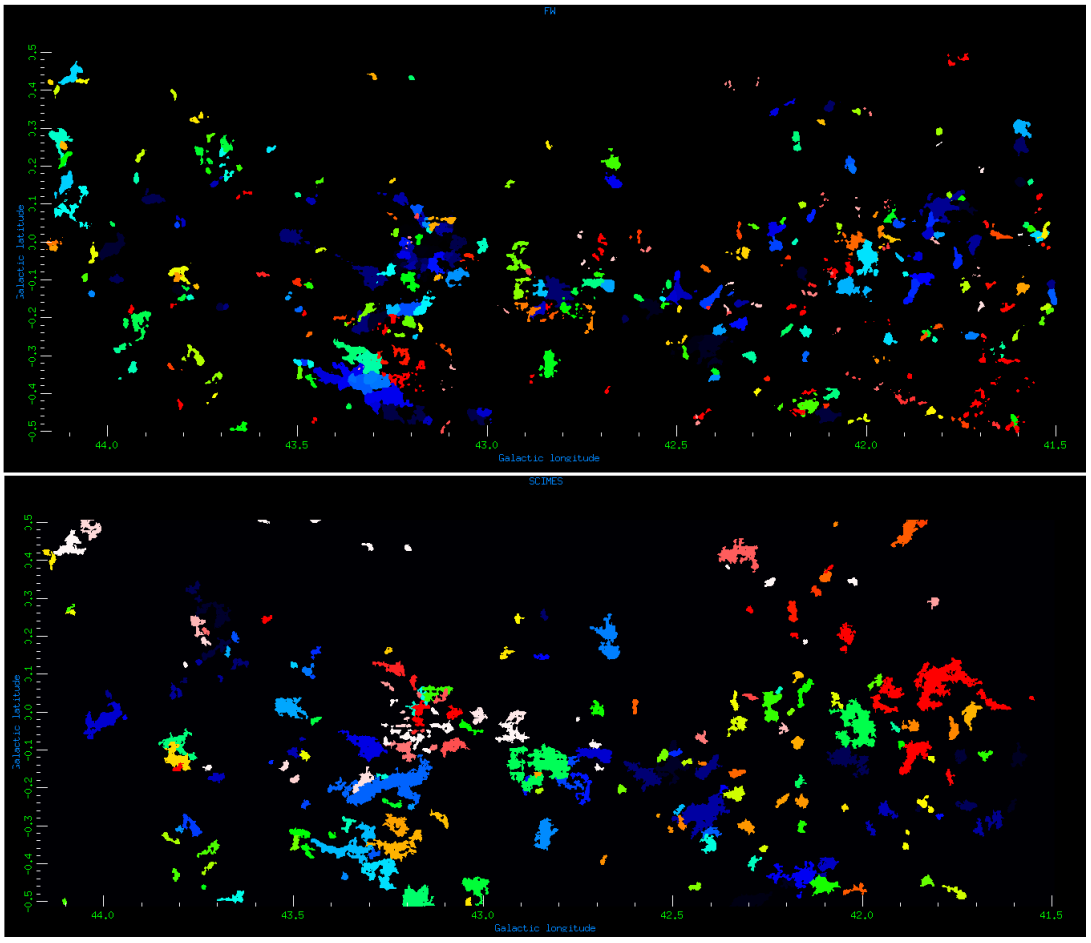


FIGURE 4.10: Projected cloud assignments in over Region 7 (see Table 4.1 in FW (top panel) and SCIMES (bottom panel). The clouds are colour-coded according their assignment numbers in FW and SCIMES. In the case of overlapping clouds, the line-of-sight projection places the cloud with the highest assignment number on top.

Region	Longitude (deg)	**FW No. clouds**	**SCIMES No. clouds**	**FW Av. size (voxels)**	**SCIMES Av. size (voxels)**	**FW Mean (voxels)**	**SCIMES Mean (voxels)**	**FW Median (voxels)**	**SCIMES Median (voxels)**
0	32.17-33.59	54	42	1223.2	416.8	1223.0	408.0	763.0	139.0
1	27.44- 31.03	945	526	1419.9	827.1	1419.0	800.0	737.0	137.0
2	30.11-33.06	925	601	1465.6	2885.6	1465.0	2671.0	732.0	275.0
3	32.82-35.17	372	314	1264.6	1441.8	1264.0	1382.0	655.5	327.5
4	34.82-37.17	289	315	1095.3	1346.8	1095.0	1241.0	518.0	347.5
5	36.82-39.84	322	306	925.5	1088.1	925.0	999.0	584.5	292.0
6	39.49-41.84	192	236	996.8	885.1	996.0	830.0	565.5	279.0
7	41.49-44.17	174	214	1083.2	780.0	1083.0	746.0	571.0	270.0
8	43.72-46.67	96	70	1891.3	642.7	1891.0	542.0	843.5	180.5
9	43.82-44.54	19	24	755.5	1383.0	755.0	1330.0	312.0	426.0

TABLE 4.1: Average size of the clouds in the FW and SCIMPS extractions over the 10 regions of CHIMPS (see also Figure 3.5). Notice that the boundary of each region has been set to the point of maximum extension, the actual cuts have irregular shapes. The table includes the average size in voxels (Av. size), the number of clouds extracted (No. clouds), the mean value (Mean) and the median (Median) of the distribution of cloud volumes in each region (Mean).

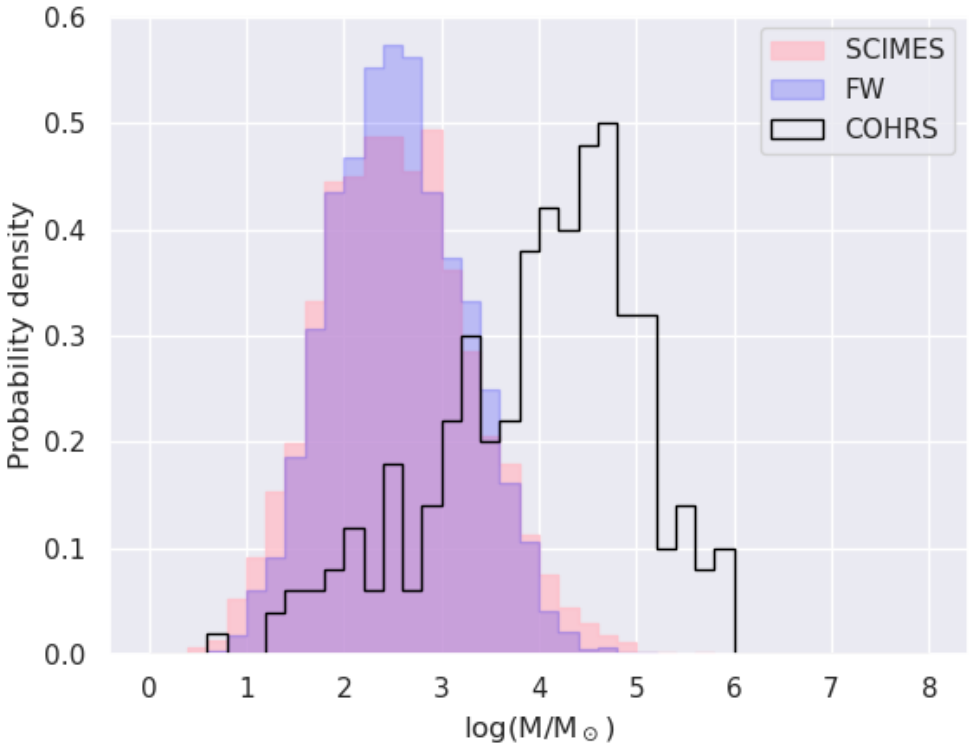


FIGURE 4.11: Distributions of masses of the CHIMPS ^{13}CO $(3-2)$ sources in the FW (blue) and SCIMES (red) catalogues. The black histogram is the distribution of ^{12}CO $3 \rightarrow 2$ sources in a subset of the COHRS catalogue (see text).

A Kolmogorov-Smirnov test reveals that SCIMES and FW volume distribution differ significantly ($k = 0.30$ with p-value $<< 0.001$.

4.4 Physical properties

4.4.1 Mass

Once distances are assigned, the true size of each voxel in the SCIMES segmentation can be calculated. Its mass and consequently the CO mass of the cloud is then estimated through the column density cubes (see section 2.1.2). The H_2 mass of the cloud is estimated by considering the mean mass per H_2 molecule, taken to be 2.72 times the mass of the proton, accounting for a helium fraction of 0.25 (Allen, 1973), and an abundance of 10^6 H_2 molecules per ^{13}CO molecule.

Figure 4.11 shows the comparison of the distributions of mass in the two CHIMPS emission extractions with the addition of the mass in COHRS. The calculation for mass estimation from CO luminosities in COHRS is explained in Colombo et al. (2019).

FIGURE 4.12: The masses associated to the CHIMPS and COHRS sources as functions of the heliocentric distance. The colours refer to the method of extraction and survey.

The mass distribution as a function of heliocentric and Galactocentric distances are presented in Figures 4.12 and 4.13 respectively. The trend at small distances in Figure 4.13 is likely to be an artefact originating from the small number of sources in the initial bin (3.5-4.0 kpc) and the position of the centre of the bin in the plot.

As expected, the larger structures detected through ^{12}CO emission result in the larger masses of Figures 4.11, 4.33, and 4.13. CHIMPS and COHRS trendlines also follow similar pattern, suggesting that the segmentation of COHRS identifies the more massive counterparts of CHIMPS obejects.

Furthermore, a Kolmogorov-Smirnov test on the FW and SCIMES mass distributions returns $k = 0.045$ with p-value $= 0.004$).

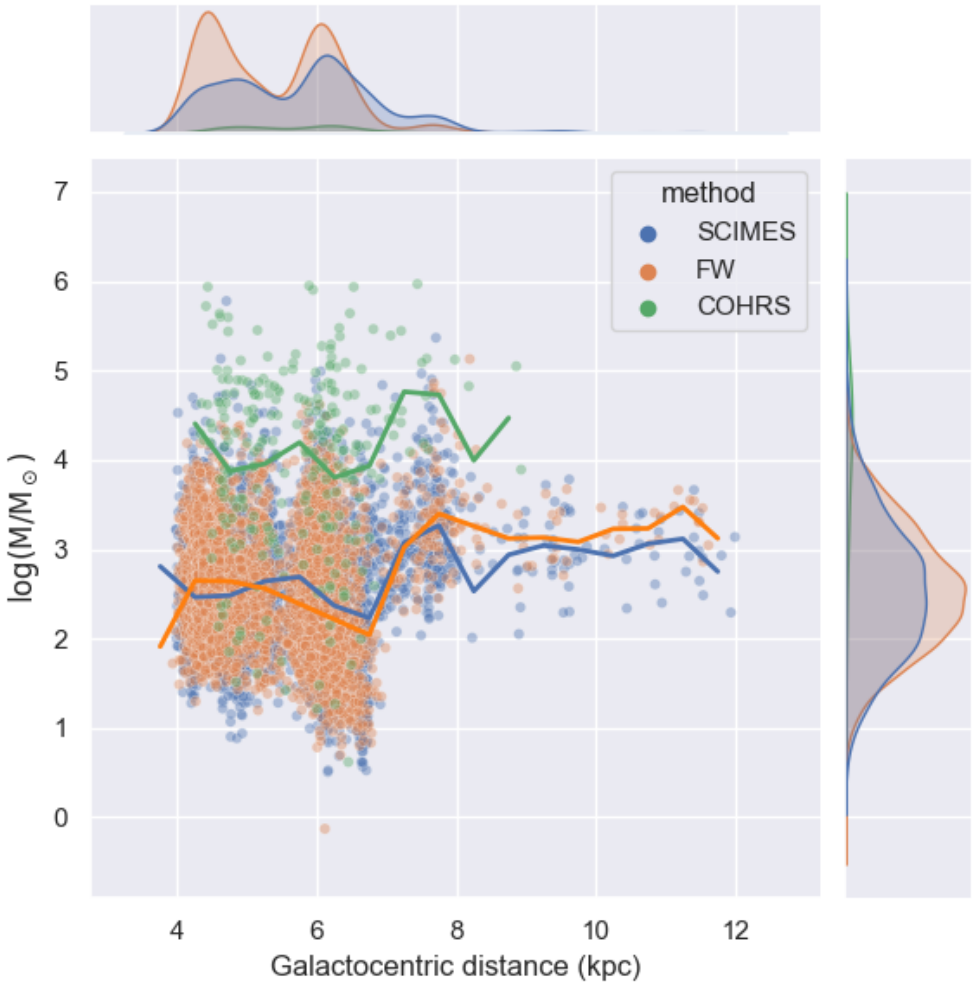


FIGURE 4.13: The masss of the CHIMPS and COHRS sources as functions of the Galactocentric distance. The trend lines show the mean values of clouds in 0.5 kpc-wide bins. The colours refer to the method of extraction and survey.

Vital to an accurate mass estimation is a precise distance assignment. The uncertainty on the distances estimated from Bayesian distance algorithm is ~ 0.3 kpc (Reid et al., 2016). This affects shorter distances the most (30% at 1 kpc) but falls to a few per cent already at 5 kpc. The other assignment methods used in the surveys are mentioned in the Chapter are described in section 2.1 and references therein. The uncertainty in the cloud mass is estimated. Taking into account the error on the conversion CO-to-H_2 conversion factor and column density estimation (Urquhart et al., 2018; Rigby et al., 2019), assuming a typical error in cloud mass of order 30-40 per cent. The uncertainties are measurement errors. In addition, the distance assignment (as well

as all other calculated parameters) is very likely to be contaminated by uncertainties in the assumptions and approximations in the variety of methods considered in the various surveys. Appendix D presents a comparison between mass distribution derived from random distance assignments, suggesting distance assignments make no significant difference to the full-sample statistics.

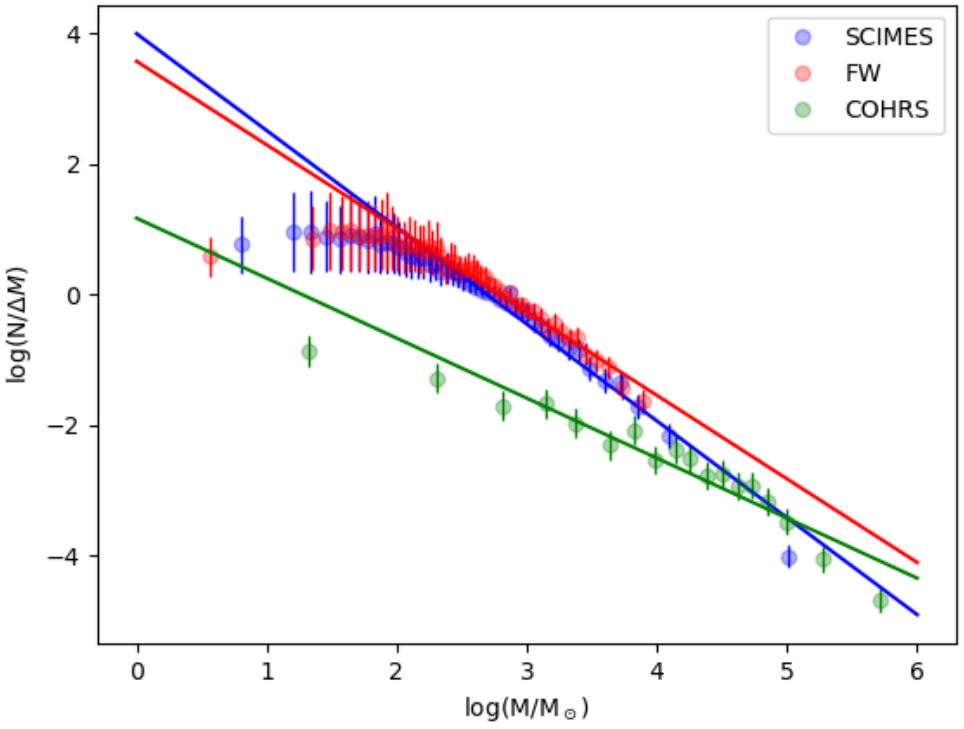


FIGURE 4.14: Comparison between the data and the fitted functions for mass spectra. The dots indicate the the centres of the mass bins. The colours refer to the method of extraction and survey.

The mass spectra for CHIMPS clouds and their fitted relations are displayed in Figure 4.14. The mass spectral indices found with a power law fit are -1.450 ± 0.029 for SCIMES clouds, -1.284 ± 0.016 for FW and -0.920 ± 0.039 for the COHRS survey. To binning of the mass follows Maíz Apellániz & Úbeda (2005) with $2N^{2/5}$ with variable width and fixed population of $2N^{2/5}$, N being the number of individuals in the entire population. This convention is adopted to remove biases due to binning. The SCIMES and FW indices are consistent with the -1.5 value found in previous studies (Sanders et al., 1985; Solomon et al., 1987; Williams et al., 1994; Roman-Duval et al., 2010).

4.4.2 Hydrogen number density

The mean (volumetric) particle density (or number density) over the approximate volume of a cloud (assuming 2D to 3D symmetry) is calculated as

$$\overline{n}(\mathrm{H_2}) = \frac{3}{4\pi}\frac{M}{\mu m_p R_{eq}^3}, \tag{4.7}$$

where M is the mass of the cloud, μm_p ($= 2.72 m_p$) is the modified proton mass.

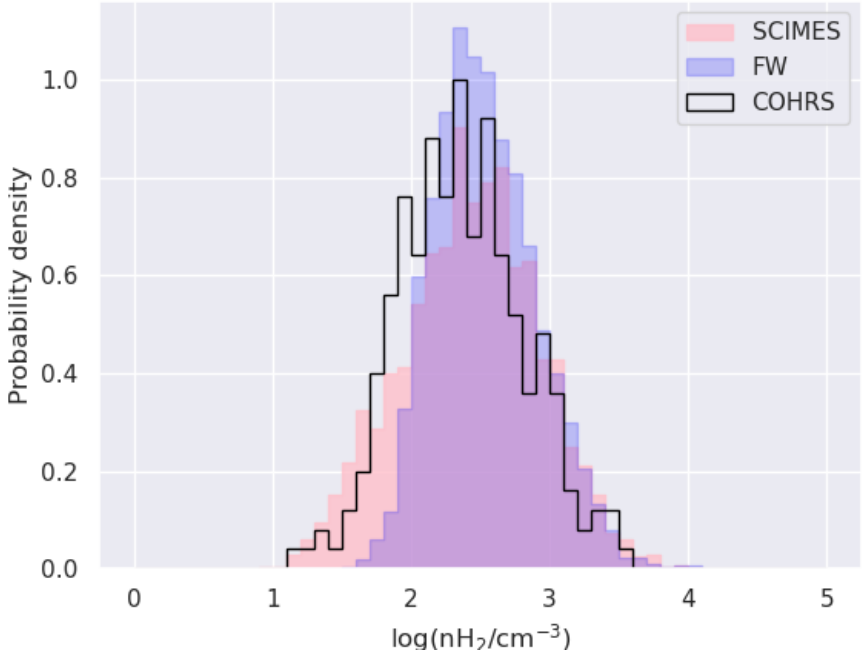


FIGURE 4.15: Distributions of the H_2 number density in the CHIMPS ^{13}CO (3 - 2) sources in the FW (blue) and SCIMES (red) catalogues. The black histogram is the distribution of the equivalent radii of the ^{12}CO (3 - 2) sources in a subset of the COHRS catalogue (see text).

The distribution of molecular hydrogen number density extracted through the FW method in CHIMPS and by SCIMES in CHIMPS and COHRS is reported in Figure 4.15, The larger masses and greater radii found in COHRS clouds result in a distribution of mean molecular hydrogen density that is comparable to the ones obtained for the SCIMES and FW segmentations.

Again, running a Kolmogorov-Smirnov test shows that the distribution of the molecular hydrogran number density in the SCIMES and FW extractions differ significantly ($k = 0.14$ with p-value $= 1.48 \times 10^{-28}$).

Figures 4.17 and 4.18 display plots of $\overline{n}(H_2)$ as a function of the heliocentric and Galactocentric distance respectively.

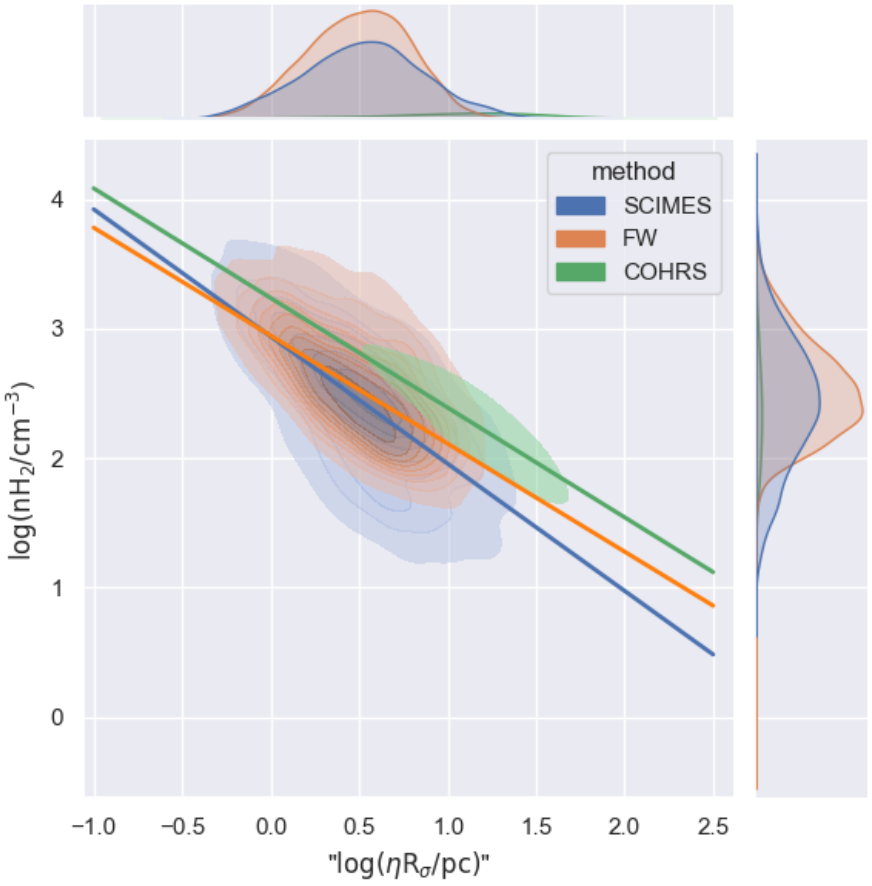


FIGURE 4.16: Size–density relationship for the CHIMPS clouds. The contour plots refer to the FW and SCIMES extractions, and a selected sample of COHRS sources (see text). The size parameter is the scaled intensity-weighted rms size (see text), ηR_σ with $\eta = 2.0$. The solid lines indicate the fitted relationships.

.

At any level of the molecular gas hierarchy, from the most compact cores of a few solar masses and densities of $\sim 10^5 \mathrm{cm}^{-3}$ to entire GMCs with mean densities of $n \sim 10^2 \mathrm{cm}^{-3}$ and masses of $10^5 - 10^6$, each identified structure usually contains many Jeans masses (Krause, 2020). The Jeans equations 1.2 and 1.1 can thus be applied to produce an estimate for the timescale for cloud collapse, and consequently for star formation within the collapsing regions. This timescale is known as the free-fall time (see below). If not

delayed by other physical mechanisms (see Chapter 1), the free-fall time depends on density ranging from ~ 3 Myr for the more rarefied regions ($n \sim 10^2 \mathrm{cm}^{-3}$) to 0.1 Myr for cores with $\sim 10^5 \mathrm{cm}^{-3}$. Furthermore, it follows from the equations 1.2 and 1.1 that, as density increases with the advancing of collapse, both Jeans length and the Jeans mass decrease. Such reduction of Jeans lengths and masses induces fragmentation in the collapsing cloud (Hoyle, 1953). Cloud fragmentation is thought to cease once the gas becomes adiabatic, which occurs at large volume densities with the gas becoming optically thick.

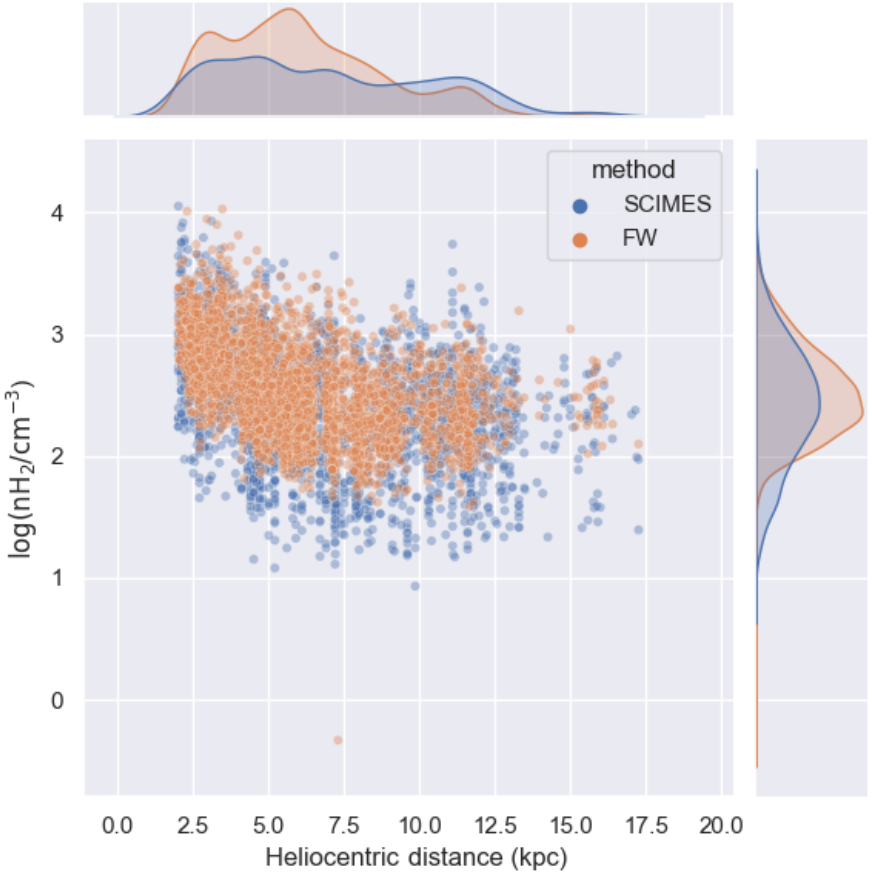


FIGURE 4.17: The H_2 number density in the CHIMPS and (a selection of) COHRS sources as functions of the heliocentric distance. The colours refer to the method of extraction and survey.

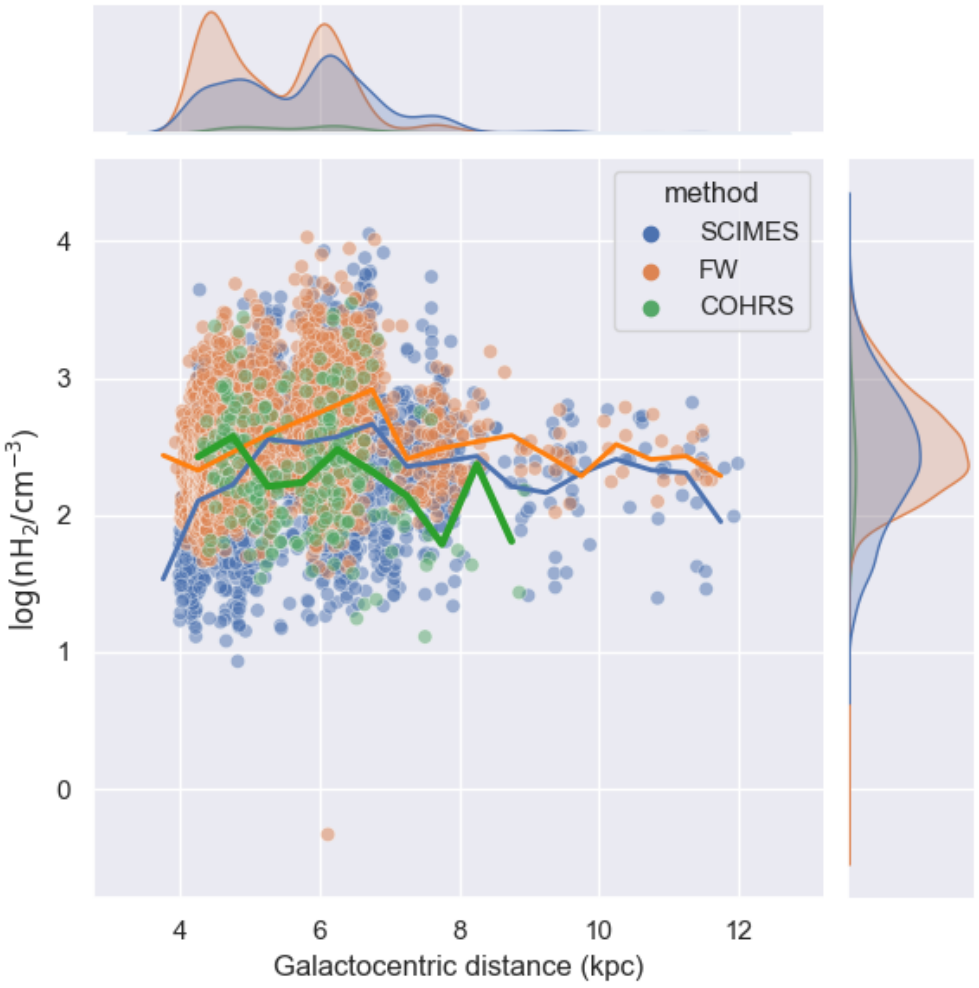


FIGURE 4.18: The H_2 number density in the CHIMPS and (a selection of) COHRS sources as functions of the heliocentric distance. The trend lines show the mean values of clouds in 0.5 kpc-wide bins. The colours refer to the method of extraction and survey.

The distribution of H_2 number densities in Figure 4.15 portrays values much less than the critical density of ^{13}CO (see Table 1.1). The H_2 number density assigned to each cloud represents the average density over the entire (approximated) volume of the cloud. This average value accounts for both clumps with a density over the critical threshold and areas of far more rarefied gas. The estimated low density from the emission segmentation is an indicator of clump formation with high-density forming clumps that may lay at scales below the telescope resolution. The volume filling factor of the gas is low at the regimes where clumps are forming. In addition, gas with densities lower than the critical density will be warmer than the calculated excitation temperature (Rigby et al., 2019). However, it may still emit in a sub-thermal mode in which the energy level populations

are not distributed according to the Boltzmann distribution. This underestimate in the gas temperature is mirrored in overestimates in the gas column density (Rigby et al., 2019). The distribution of mean excitation temperatures of the FW extraction of CHIMPS clouds is found to have a mean value of 11.5 K, which matches the expectation for molecular structures covering the size regime from cores, through clumps, to clouds (Bergin & Tafalla, 2007). Sub-thermal emission can therefore be assumed not to be a dominant effect here.

Applying a power-law fit to the size-density relation shown in Figure 4.16, produces average number densities proportional to R^a with $a = -0.982 \pm 0.004$ for SCIMES clouds, and $a = -0.834 \pm 0.007$ in the FW case. The FW value departs significantly from the original scaling relation $a = -1.1 \pm 0.005$ found by Larson (1981) indicating that the smallest CHIMPS cloud are less dense than would be predicted by the Larson relationship. For COHRS clouds $a = -0.846 \pm 0.067$.

4.4.3 Free-fall and crossing times

The free-fall timescale, $t_{\rm ff}$, represents the characteristic time that would take a body to collapse under its own gravitational attraction. As mentioned above, $t_{\rm ff}$ depends solely on the density and the chemical species of the gas. In terms of the molecular hydrogen mean number density discussed in the previous sub-section,

$$t_{\rm ff} = \sqrt{\frac{3\pi}{32 G \mu_{mp} \overline{n}(H_2)}}. \tag{4.8}$$

The crossing timescale, $t_{\rm cross}$, corresponds to the time it takes a disturbance to cross the system at the sound/signal speed in the medium. The length of $t_{\rm cross}$ is directly proportional to the size of the system and inversely proportional to the velocity dispersion (defined by equation 4.10) of the gas:

$$t_{\rm cross} = \frac{2R_{\rm eq}}{\sigma_v}. \tag{4.9}$$

The distributions of these timescales for the two segmentations of CHIMPS and COHRS are compared in Figures 4.19 and 4.20. Kolmogorov-Smirnov tests on both the crossing

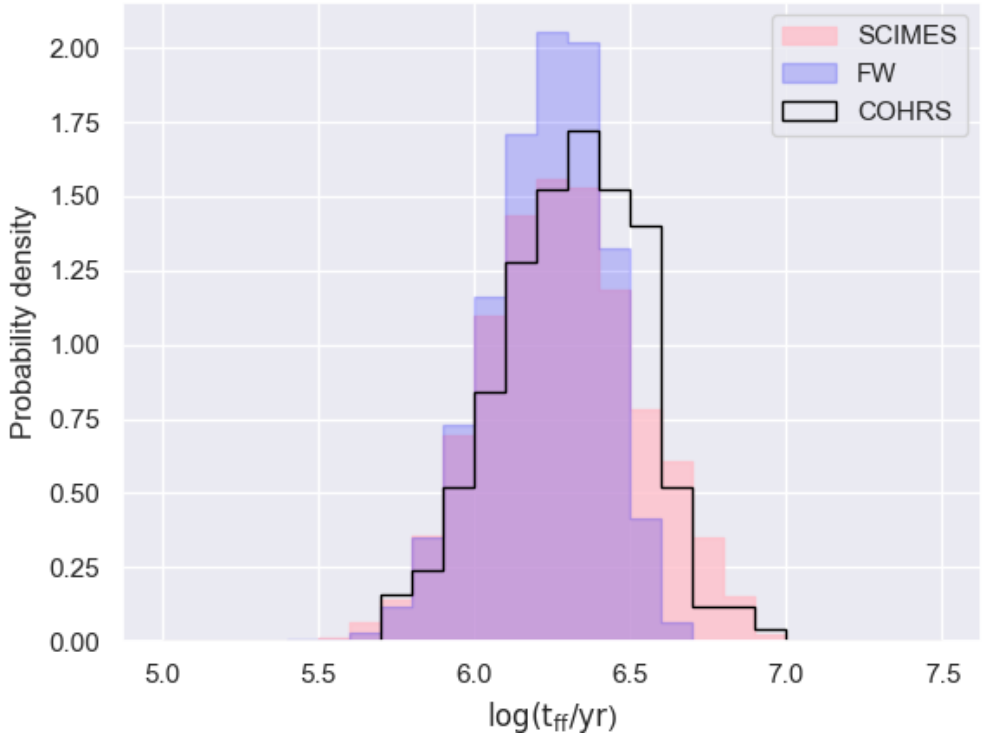


FIGURE 4.19: Distributions of the free fall time associated with the CHIMPS ^{13}CO (3 - 2) sources in the FW (blue), SCIMES (red), and COHRS (black) catalogues.

time and the free-fall time distributions of the SCIMES and FW clouds shows that the (null) hypothesis that both distributions are two samples of the same distribuition cannot be accepted (free-fall time $k = 0.14$ with p-value $= 5.42 \cdot 10^{-28}$, crossing time $k = 0.40$ with p-value $<< 0.001$.

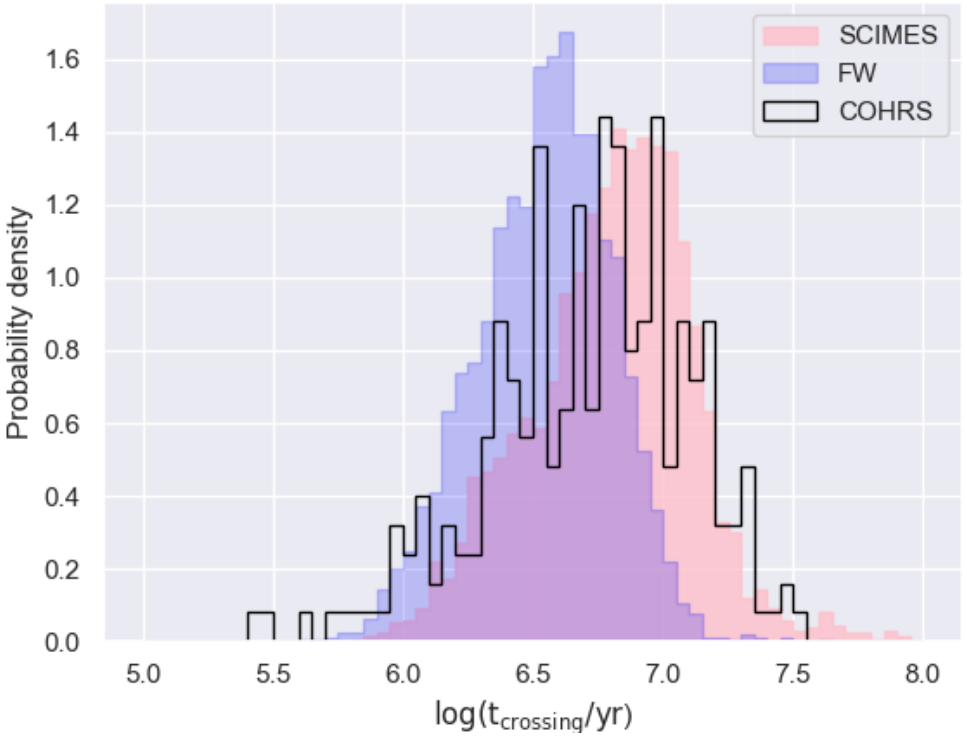


FIGURE 4.20: Distributions of the crossing time associated with the CHIMPS ^{13}CO (3 - 2) sources in the FW (blue) and SCIMES (red), and COHRS (black) catalogues.

4.4.4 Velocity dispersion

The velocity dispersion of gas in molecular clouds is measured as the intensity-weighted rms deviation of voxels from the centroid in the spectral direction (Berry, 2015):

$$\sigma_v = \sqrt{\frac{\sum d_i v_i^2}{\sum d_i} - \left(\frac{\sum d_i v_i}{\sum d_i}\right)^2} \tag{4.10}$$

where d_i is the data value at the voxels i The summations are intended over all voxels in a cloud. This definition is equivalent to using the intensity-weighted second moment of velocity [5].

[5]See dendrogram statistic `https://dendrograms.readthedocs.io/en/stable/catalog.html` and cube moments defined in `https://spectral-cube.readthedocs.io/en/latest/api/spectral_cube.SpectralCube.html#spectral_cube.SpectralCube.moment`.

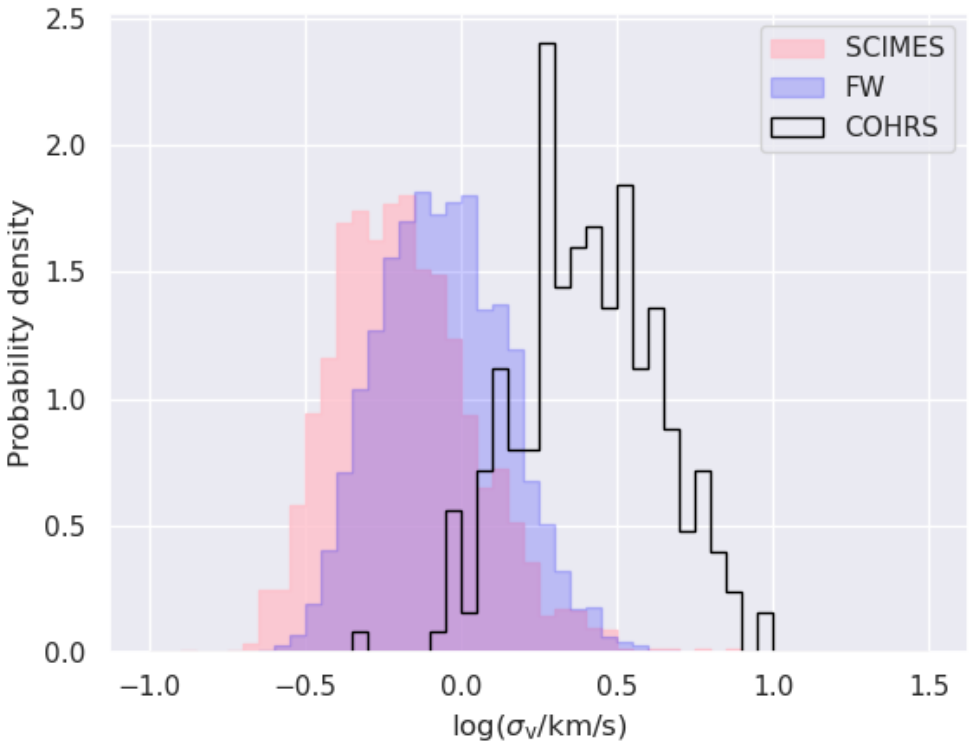


FIGURE 4.21: Distributions of the velocity dispersion in the CHIMPS ^{13}CO $(3-2)$ sources in the FW (blue) and SCIMES (red) catalogues.

For a cloud with a Gaussian distribution of velocities, the velocity dispersion equals the standard deviation of the Gaussian. In general, the larger the size of a cloud, the wider the distribution of velocities of its particles, thus its velocity dispersion. The velocity dispersion causes the broadening of linewidth in CO observation. This fact is mirrored in the distribution of velocity dispersions in the clouds of the COHRS catalogue (Figure 4.21) and their size-linewidth relation in Figure 4.22. Line widths are expected to be larger in ^{12}CO because of the high optical depths suppressing the peak intensities as well as tracing larger structures with larger turbulent velocities. Applying a power-law fit to the size-velocity dispersion relation shown in Figure 4.22, produces $\sigma_v \propto R^a$ with $a = 0.310 \pm 0.004$ for SCIMES clouds, and $a = 0.341 \pm 0.003$ in the FW case. Both values are similar to the original scaling relation $a = 0.38$ found by Larson (1981) over a factor of 30 in size, which was originally interpreted as evidence that the internal motions of molecular clouds follow a continuum of turbulent flow inherited from the ISM at larger scales. For the COHRS clouds $a = 0.277 \pm 0.011$.

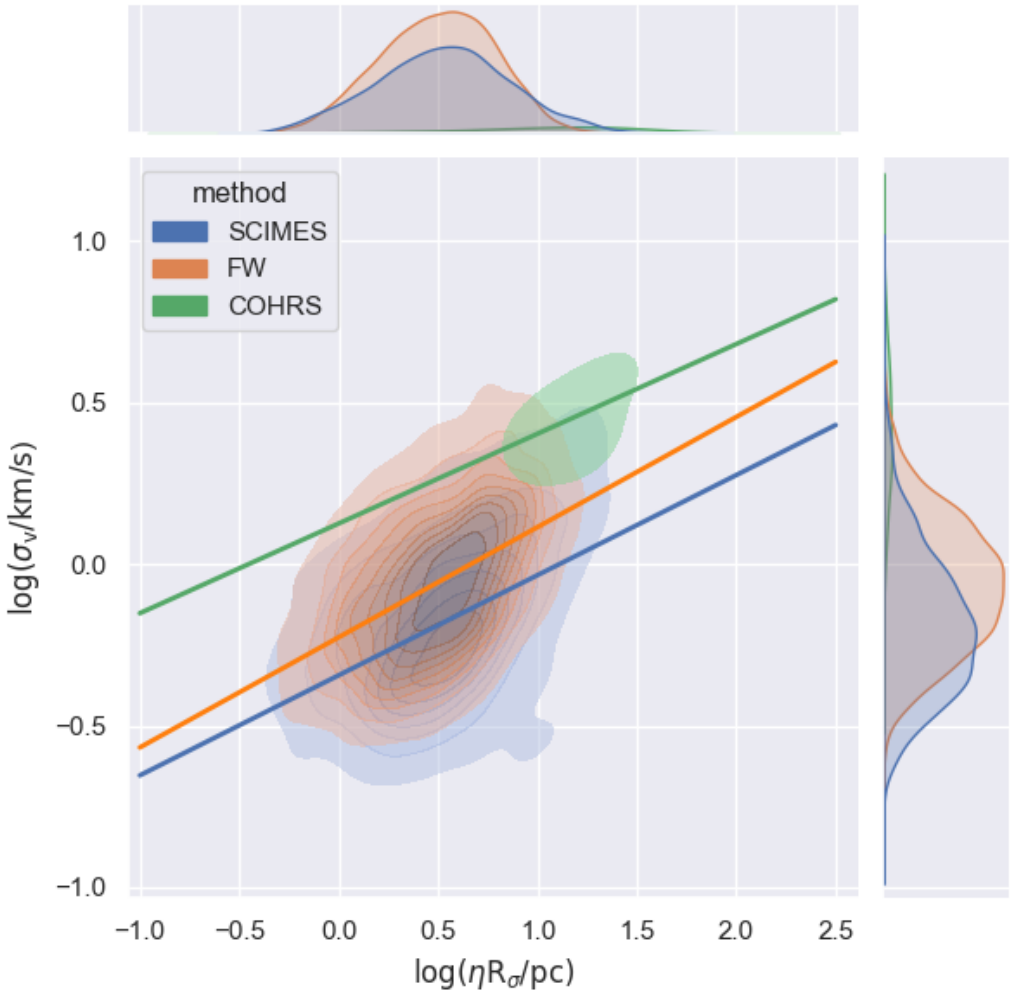


FIGURE 4.22: Size–linewidth relationship for the CHIMPS clouds. The contour plots refer to the FW and SCIMES extractions. The size parameter is the scaled intensity-weighted rms size (see text), ηR_σ with $\eta = 2.0$. The solid lines indicate the fitted relationships.

.

A Kolmogorov-Smirnov test show SCIMES and FW distribution in Figure 4.21 are significantly different and cannot be identified as two samples of the same distribution ($k = 0.25$ with p-value $= 0.001$.

4.4.5 Excitation temperature

Excitation temperatures are assigned to clouds through masking of the temperature maps constructed in section 2.1.2. These data cubes were constructed by Rigby et al.

(2019) and the temperature assignment used in this section follows the method described in their article. A unique excitation temperature is then assigned to each cloud by taking emission weighted average of the temperatures of its voxels.

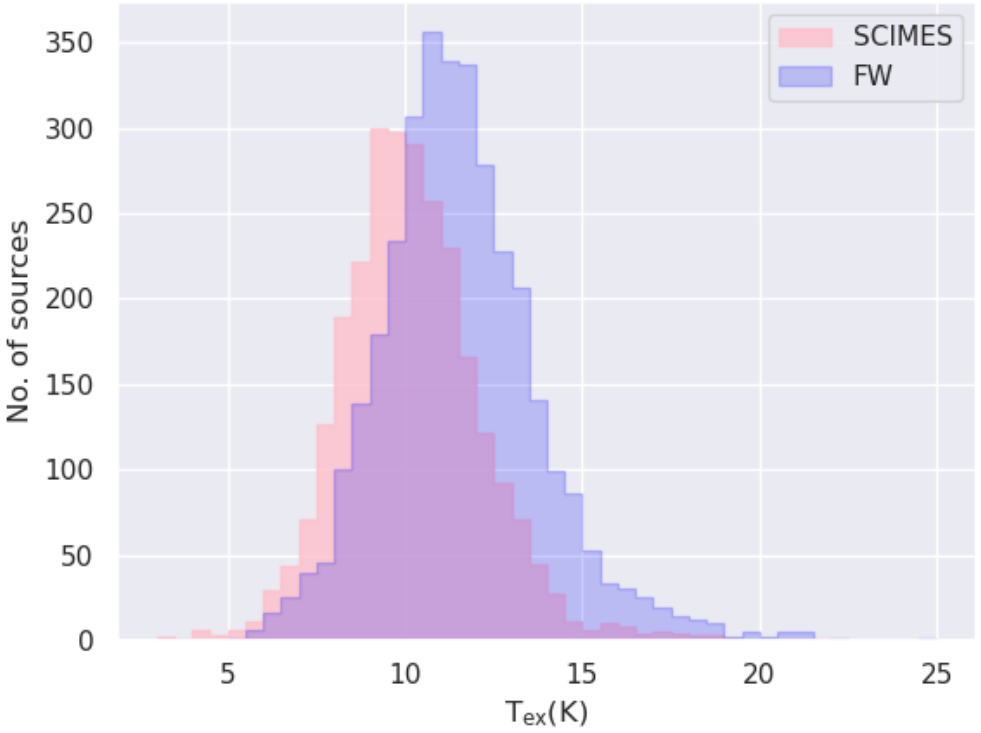


FIGURE 4.23: Distribution of excitation temperatures in CHIMPS. The colours refer to the method of extraction.

The distributions of excitation temperature in the FW and SCIMES segmentations of the ^{13}CO (3-2) emission in CHIMPS are shown in Figure 4.23. Excitation temperatures do not vary significantly with distances (Figures 4.24 and 4.25), with the temperatures from the SCIMES catalogue being everywhere lower than FW temperatures. The average SCIMES excitation temperature is 10.19 K while FW clouds have a mean of 11.54 K. Applying a Kolmogorov-Smirnov test to the SCIMES and FW distributions of the excitation temperature shows that they cannot be characterised as two samples of the same distribution ($k = 0.28$ with p-value $<< 0.001$).

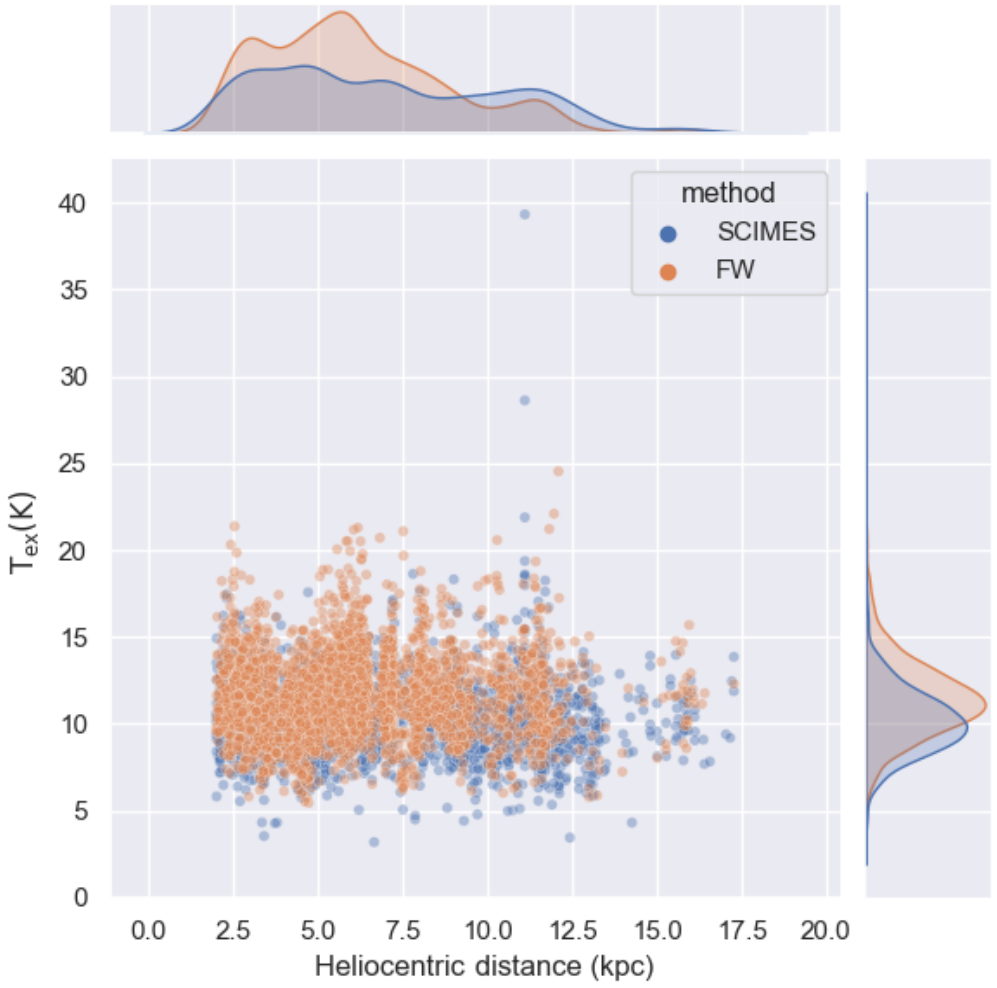


FIGURE 4.24: The excitation temperature associated weight the CHIMPS and COHRS sources as functions of the Galactocentric distance. The colours refer to the method of extraction and survey.

As a function of the Galalctocentric distance (Figure 4.25, the the two segmentations show no obvious (difference in) biases and no gradient of the excitation temperature. This contrasts the probable gradient in stellar radiation field, dominated by cosmic-ray heating or (less likely) by internal heating. Arm radii only see an increase in source counts, which in turn increases the detected scatter to higher Tex, but does not results in a significant change in the mean.

The high-temperature outliers in the SCIMES segmentation have coordinates and distances compatible with those of the star-forming region W49 ($l \approx 43.2°$, $b \approx 0.0°$ at 11.1 kpc).

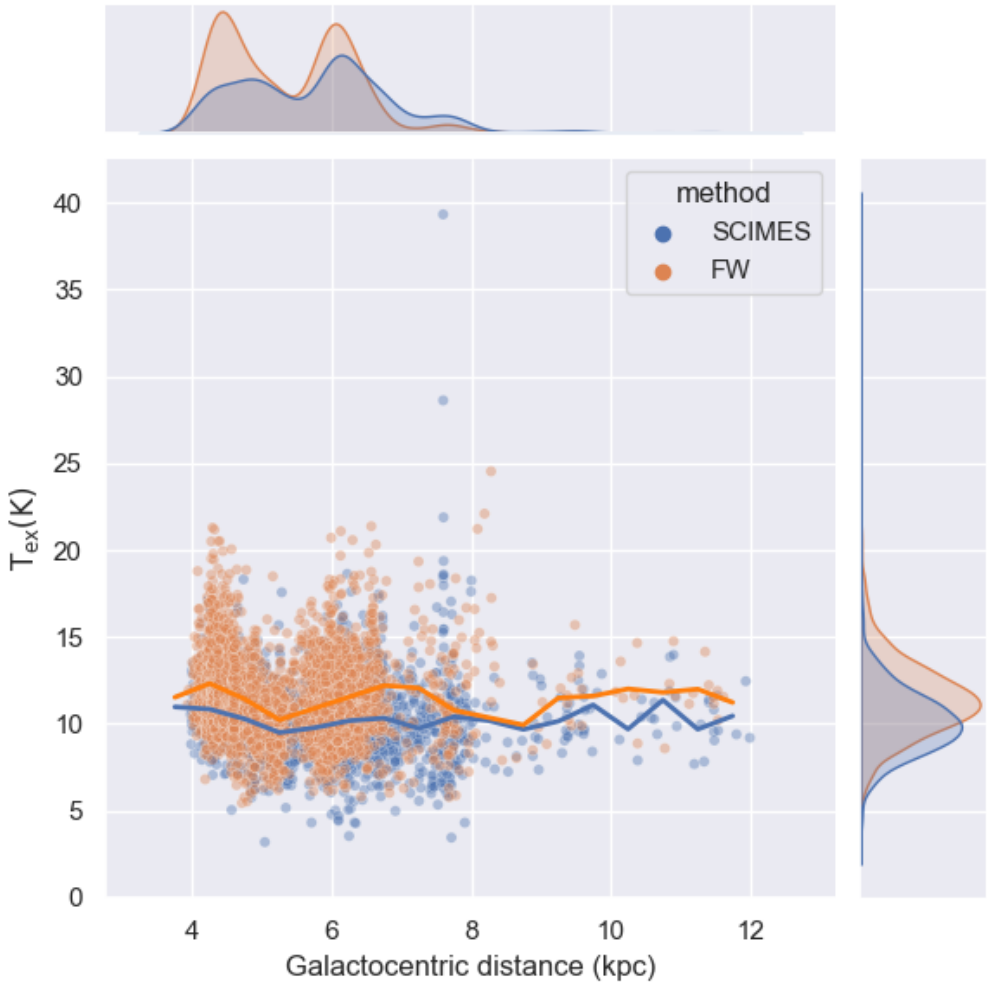


FIGURE 4.25: The excitation temperature associated the CHIMPS and COHRS sources as functions of the Galactocentric distance. The trend lines show the mean values of clouds in 0.5 kpc-wide bins. The colours refer to the method of extraction.

4.4.6 Turbulent pressure

The three-dimensional velocity dispersion ($3\sigma_v^2$) can be decomposed into its thermal

$$\sigma_T^2 = k_B T\mathrm{ex}/\mu m_p \tag{4.11}$$

and non-thermal (turbulent)

$$\sigma_{\mathrm{NT}}^2 = 3\sigma_v^2 - \sigma_T^2 \tag{4.12}$$

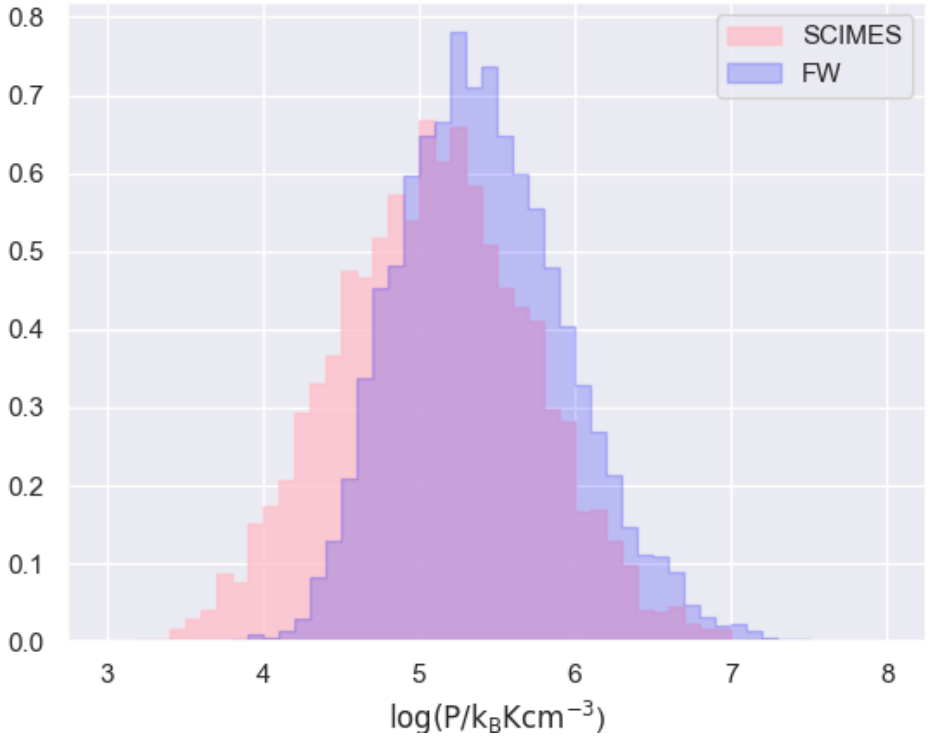


FIGURE 4.26: Distributions of the turbulent pressure associated with the CHIMPS ^{13}CO (3 - 2) sources in the FW (blue) and SCIMES (red) catalogues.

components, where the one-dimensional velocity dispersion is defined in sub-section 4.4.4. As usual, $m_{13_{CO}}$ is the mass of the ^{13}CO isotopologue and k_B the Boltzmann constant.

The turbulent pressure is then defined as

$$P_{\rm turb}/k_B = \mu m_p \overline{n}(H_2)\sigma^2_{\rm NT}/k_B, \tag{4.13}$$

$P_{\rm turb}/k_B$ has units of $K/{\rm cm}^3$.

The turbulent pressure distributions in Figure 4.26 show that SCIMES sources tend to have lower pressure then their FW counterparts. The median values of the two distributions are comparable with SCIMES having a median of 2.5×10^5 $K/{\rm cm}^3$ and FW of 4×10^5 ${\rm cm}^3$. Both these values agree with the total mid-plane pressure in the Solar neighbourhood ($\sim 10^5$ $K/{\rm cm}^3$).

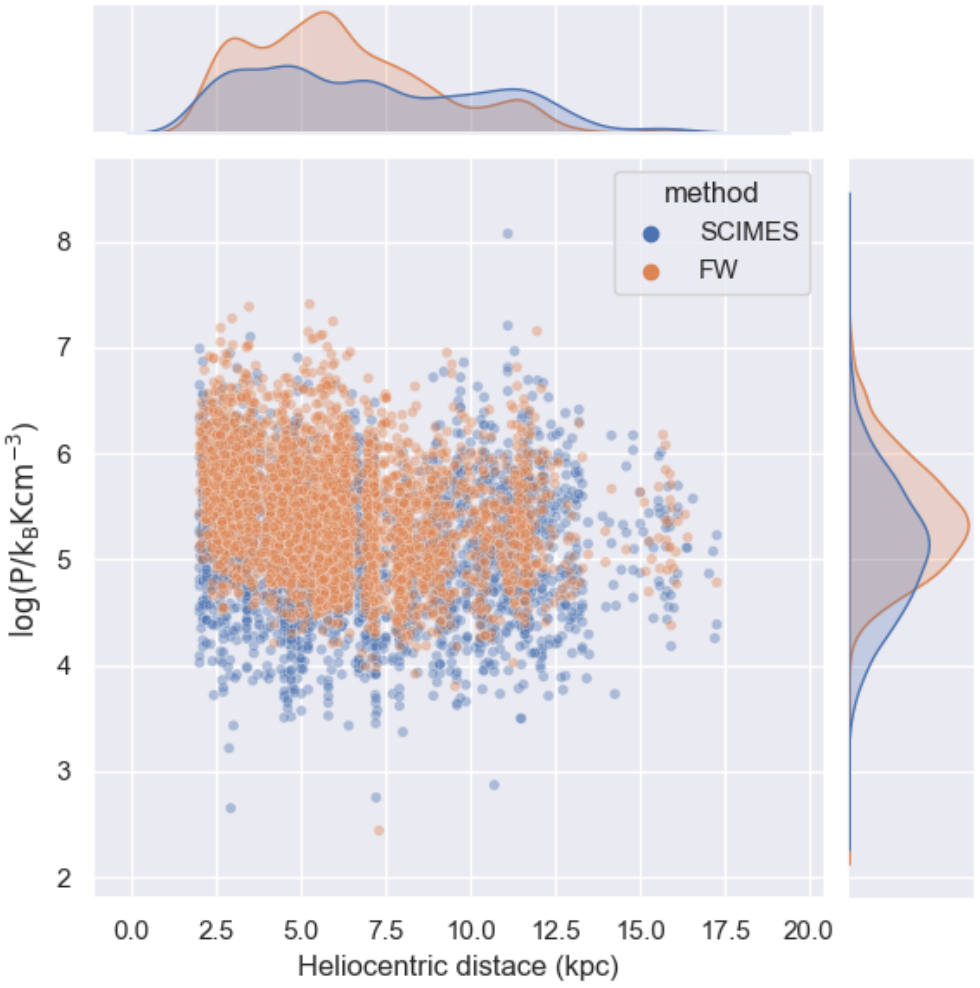


FIGURE 4.27: The turbulent pressure associated the CHIMPS sources as a function of the heliocentric distance. The colours refer to the method of extraction.

To check if the FW and SCIMES distribution of turbulent pressures differ significantly, a Kolmogorov-Smirnov test is performed. The test yields $k = 0.231$ with p-value $<< 0.001$ establishing that the null hypothesis of the two samples being drawn from the same distribution can be rejected.

The distribution of $P_{\rm turb}/k_B$ with helio- and Galactocentric distance are given in Figures 4.27 and 4.28 respectively. The range of $P_{\rm turb}/k_B$ covered by both distributions is consistent with the mid-plane values (Rathborne et al., 2014).

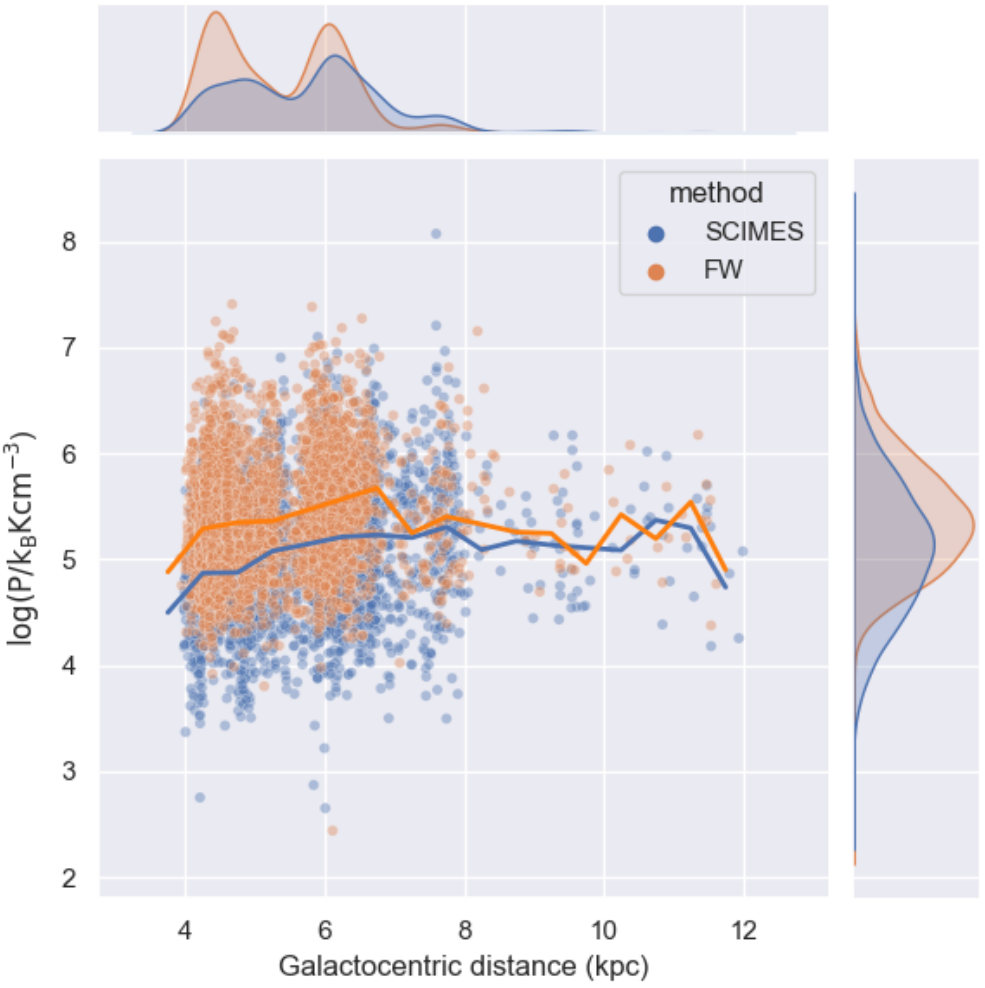


FIGURE 4.28: The turbulent pressure associated the CHIMPS sources as a function of the Galactocentric distance. The trend lines show the mean values of clouds in 0.5 kpc-wide bins. The colours refer to the method of extraction.

The thermal pressure can be defined as

$$P_{\text{thermal}} = \overline{n}(H_2) k_B T_{\text{ex}}. \tag{4.14}$$

Thermal pressure distributions are presented in Figure 4.29. The turbulent pressures are found to be ~ 60 times greater than the corresponding thermal pressures. Lower average densities result in lower pressures associated with the COHRS sample. A Kolmogorov-Smirnov test performed on the FW and SCIMES thermal pressure distributions returns $k = 0.15$ with p-value $<< 0.001$.

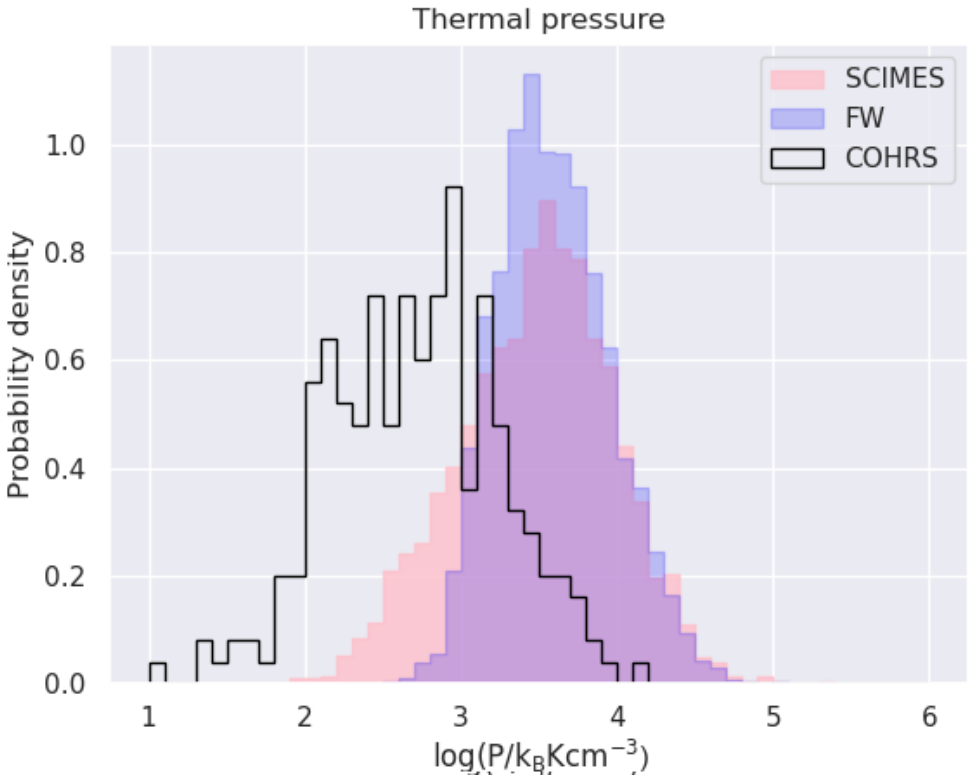


FIGURE 4.29: Distribution of the thermal pressure associated to the CHIMPS ^{13}CO (3 - 2) sources in the FW (blue) and SCIMES (red) catalogues.

4.4.7 Mach numbers

The Mach number is a dimensionless quantity that describes the dynamic state of a flow of a fluid representing the ratio of flow velocity and the local speed of sound in the medium considered. The definition of Mach number can be recast as in terms of the thermal and non-thermal components of the velocity dispersion defined above:

$$M = \sigma_{\rm NT}/\sigma_T. \tag{4.15}$$

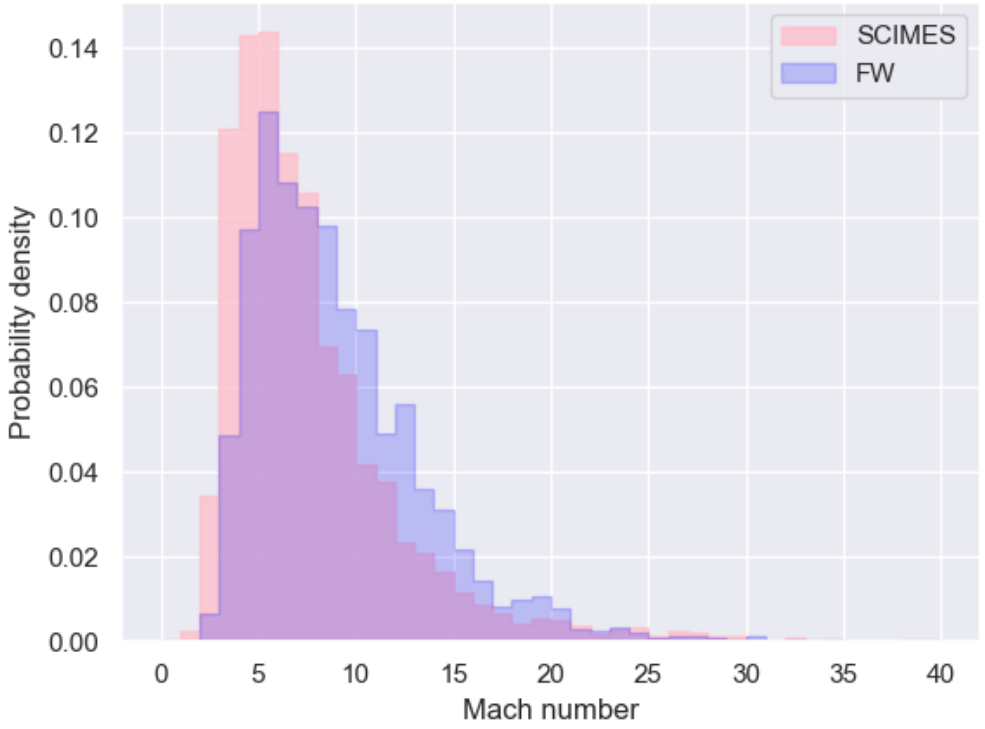


FIGURE 4.30: Distributions of the Mach numbers associated with the CHIMPS ^{13}CO (3 - 2) sources in the FW (blue) and SCIMES (red) catalogues.

Figures 4.30 represents the distributions of Mach numbers of the sources in the FW and SCIMES segmentations. The distributions look similar, both peaking in the supersonic regime ($M \sim 5$) and extending out to higher Mach numbers. The flow of molecular gas in clouds (a characteristic of turbulence) is linked to velocity dispersion, which in turn is linked to the size of the cloud. To check if the FW and SCIMES distribution of Mach number differ significantly, a Kolmogorov-Smirnov test is performed. The test yields $k = 0.21$ with p-value $<< 0.001$ establishing that the null hypothesis of the two samples being drawn from the same distribution can be rejected.

The higher number of (larger) clouds in FW results in its shift towards a higher Mach number. The difference in the distributions vanishes as the tails of the distributions flatten out past $M = 20$ where fewer large enough clouds to sustain these regimes are found.

4.4.8 The virial parameter

The virial parameter encodes the dynamic state of a molecular cloud, assuming that the cloud is capable of sustaining virial equilibrium, i.e. the virial theorem holds for the cloud, its gravitational energy Ω equals twice the kinetic energy K

$$\Omega = -2K. \tag{4.16}$$

The virial parameter is defined as the ratio of a cloud's spherically symmetric virial mass to its total mass (M)

$$\alpha_{\rm vir} = \frac{3\sigma_v^2 \eta R_\sigma}{GM} \tag{4.17}$$

where G is the gravitational constant.

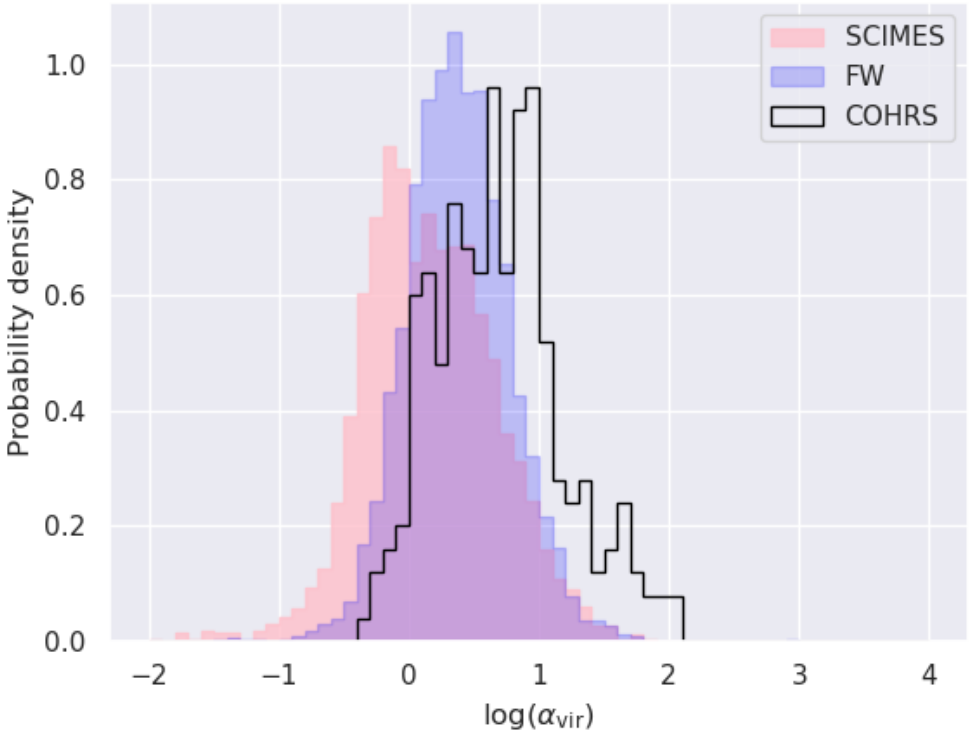


FIGURE 4.31: Distributions of the virial parameter associated with the CHIMPS ^{13}CO ($3 \rightarrow 2$) sources in the FW (blue) and SCIMES (red) catalogues. The black histogram is the distribution of ^{12}CO ($3 \rightarrow 2$) sources in a subset of the COHRS (see text).

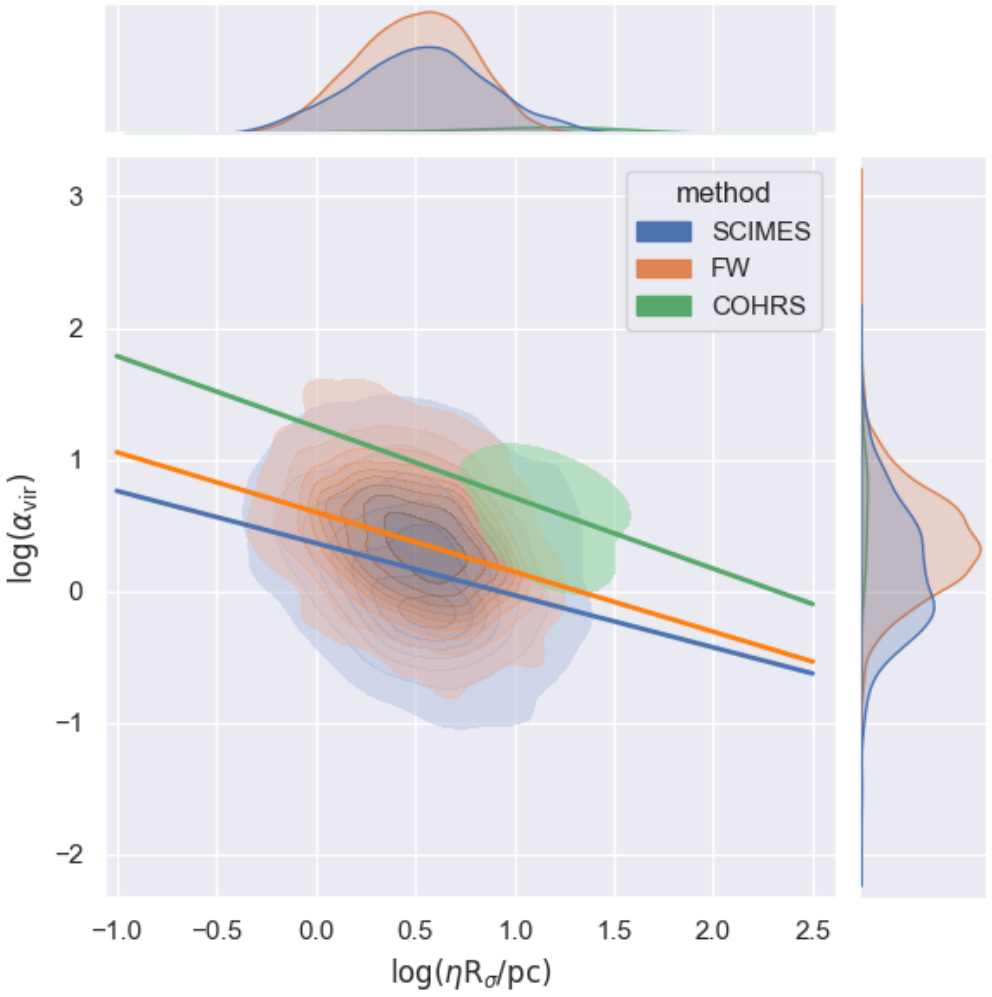


FIGURE 4.32: The relationship between the source size and the virial parameter for the CHIMPS and COHRS clouds. The contour plots refer to the FW and SCIMES extractions and the reduced fiducial sample of COHRS. The size parameter is the scaled intensity-weighted rms size (see text), ηR_σ with $\eta = 2.0$. The solid lines indicate the fitted relationships.

The definition above was given in Rigby et al. (2019) and assumes the cloud is spherical and has a radial density distribution e.g. $\rho(r) \propto r^{-2}$ (MacLaren et al., 1988). Notice that the definition includes R_σ to account for the median emission profile. The intensity-weighted radius constitutes a weighting system for gravitational energy. This weighting reinforces the gravitational energy in those regions where the density is higher. In addition, R_σ is less affected by variations in S/N levels.

Approximating a source as a spherically symmetric distribution of density introduces a factor-two uncertainty on the estimation of the virial parameter. This arises from both characterising the source by a single radius and choosing this particular radial profile.

Caution should thus be exercised in the interpretation of results involving measurements of the virial parameter.

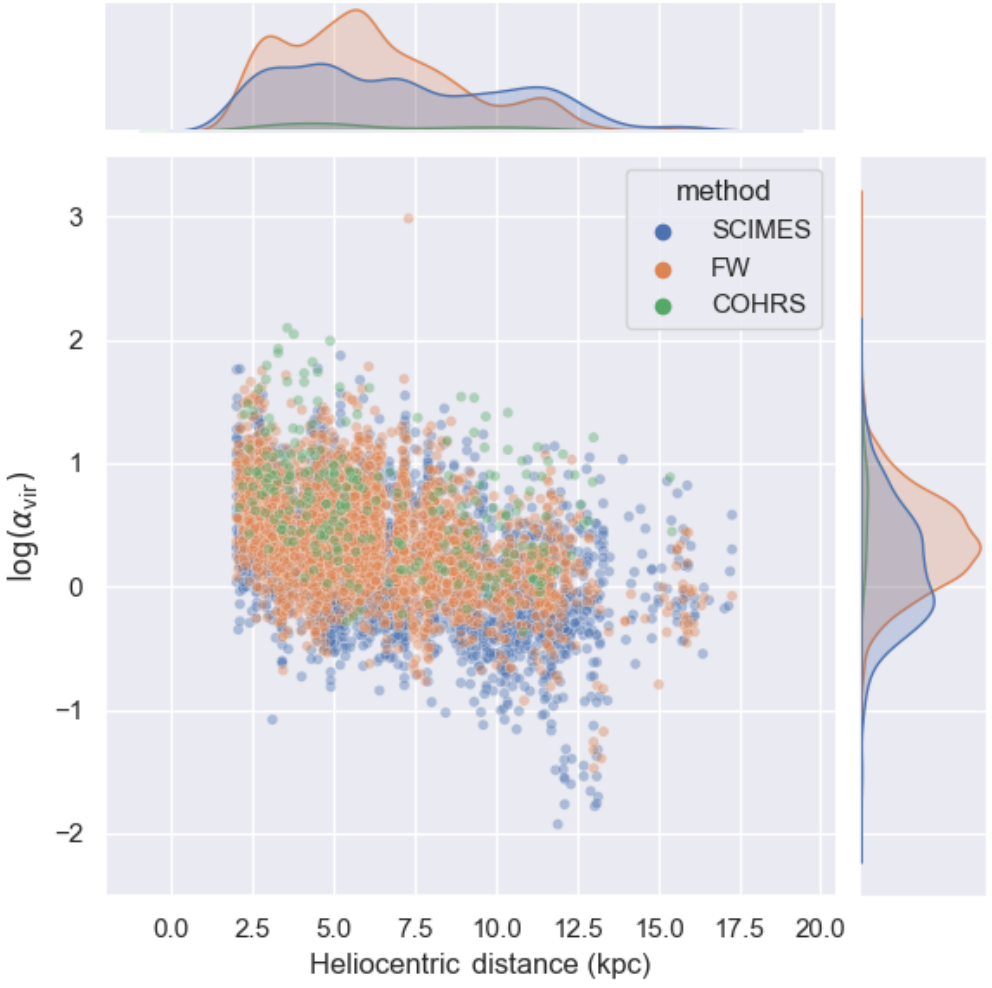


FIGURE 4.33: The virial parameter associated with the CHIMPS and (a selection of) COHRS sources as functions of the heliocentric distance. The colours refer to the method of extraction and survey. The colours refer to the method of extraction and survey.

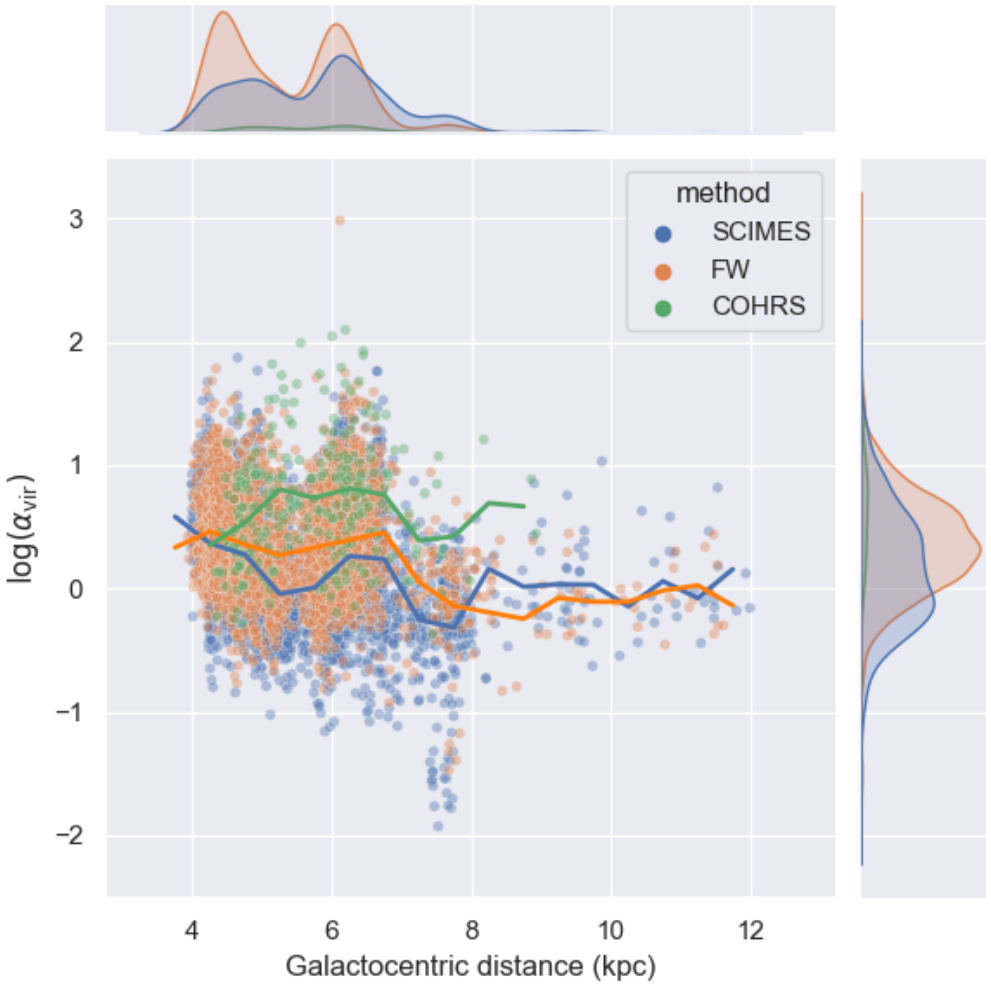


FIGURE 4.34: The virial parameter associated the CHIMPS and COHRS sources as functions of the Galactocentric distance. The trend lines show the mean values of clouds in 0.5 kpc-wide bins. The colours refer to the method of extraction and survey.

In the absence of a strong magnetic field, $\alpha_{\rm vir}$ equals 1 when the clouds are in virial equilibrium. A value $\alpha_{\rm vir} = 2$ indicates that the gravitational energy equals the kinetic energy in the cloud. Values of $\alpha_{\rm vir}$ smaller than 1 characterise an unstable, collapsing system (when other sources of supporting pressure are absent). A dissipating system, dominated by kinetic energy, is characterised by $\alpha_{\rm vir} > 2$. While $1 < \alpha_{\rm vir} < 2$ indicates approximate equilibrium. It has also been suggested that the heightened velocity dispersions due to rapidly infalling gas in collapsing cloud fragments may still raise the cloud's value of the virial parameter to ~ 2 (Kauffmann et al., 2013a). Fragments with $\alpha_{\rm vir} << 2$ are more likely to host and be supported by strong magnetic fields or to house ongoing high-mass star formation (Kauffmann et al., 2013a).

The distribution of the virial parameter in CHIMPS and COHRS is presented in Figure 4.31, while Figure 4.32 illustrates the relation between the virial parameter and cloud's size. The SCIMES distribution indicated that a large number of clouds in this segmentation are collapsing or in approximated equilibrium. SCIMES clouds are distinguished by smaller values of the virial parameter (> 0.6) fall in a size range between 2 and 20 pc, thus including the smallest, most compact sources, likely sites of star formation.

Applying a power-law fit to the size-virial parameter dispersion relation shown in Figure 4.32, produces $\alpha_v \propto R^a$ with $^-0.396 \pm 0.009$ for SCIMES clouds, and $a = -0.454 \pm 0.006$ in the FW case. Both values are significantly lower than the original scaling relation $a = -0.14$ found by Larson (1981). The observed discrepancy may be due to the varying mass completeness as a function of distance. A factor $^-0.538 \pm 0.026$ was found for COHRS clouds.

Performing a Kolmogorov-Smirnov test on the FW and SCIMES distributions of the virial parameter reveals these two distributions differ significanly. The null hypothesis of the two samples being drawn from the same distribution must thus be rejected ($k = 0.26$ with p-value $<< 0.001$).

Figures 4.33 and 4.34 show the virial parameter as a function of the Heliocentric and Galactocentric distances respectively. A closer look to the trendlines in Figure 4.34 reveals hint of a slightly increased $\alpha_{\rm vir}$ inside 7 kpc, or perhaps in the spiral arms. This trend may be due to the errors on the means of the bins increasing significantly at large radii.

4.5 Summary

This chapter presents an attempt to cross-correlate the physical properties of the molecular clouds extracted from CHIMPS ^{13}CO (3-2) emission maps through the FW and SCIMES algorithms. These methods produce different numbers of molecular clouds (SCIMES 2944, FW 3665), with similar ranges in masses, volumes (number of voxels), equivalent radii mean number densities, and velocity dispersions. SCIMES produces slightly wider ranges of sizes (volumes and equivalent radii) which suggests that the size and number of clouds extracted may both depend of algorithmic paradigm and the Galactic environment (see Chapter 7). The distributions of mean number densities,

masses, the virial parameters, and dynamic timescales mirror the differences in volumes and geometries found in the two segmentations. The distributions of velocity dispersions only depend on the size of the clouds as identified by each algorithm.

Chapter 5

Analysis of turbulence: Methods

The study of the structure of giant molecular clouds relies upon obtaining information on the three-dimensional distribution of significant physical fields, such as density, temperature, and velocity. In practice, however, the description of these fields in position-position-velocity datasets is limited to their projection along the line of sight onto a two-dimensional spatial coordinate plane and a spectral component. To retrieve the intrinsic properties of a three-dimensional field from its projected two-dimensional counterpart is a complicated task. An obvious example is the derivation of the three-dimensional volume density distribution from the observed (projected) column density.

In recent years, Brunt et al. (2010) and Brunt & Federrath (2014) developed a method to overcome these complications and reconstruct specific properties of the original field from limited observational information. Their method relies on the properties of Fourier transforms and symmetry arguments to recover the averaged properties of the full three-dimensional field from the projected observables. This method is particularly well-suited for the study of turbulent motions within velocity and momentum fields (see below) for which only the line-of-sight component can be measured (Brunt et al., 2010). In this case, the line-of-sight component splits naturally into a solenoidal (divergence-free) and a compressive (curl-free) component through a Helmholtz decomposition. Solenoidal and compressive modes of turbulence are believed to be associated with the star-formation efficiency in molecular clouds. In this framework, the high star formation efficiency

(SFE) observed in spiral-arm clouds is linked to the prevalence of compressive turbulent modes. In contrast, the low SFE that characterises clouds in the Central Molecular Zone (CMZ) is related to the shear-driven solenoidal component. The key quantity for studying the SFE in terms of turbulent mode is the solenoidal fraction which encodes the relative amount of power in the solenoidal modes of the momentum density field characterising a molecular cloud.

5.1 The solenoidal fraction

The method developed by Brunt et al. (2010); Brunt & Federrath (2014) allows us to quantify the relative fraction of the solenoidal and compressive turbulence modes. This section presents the main concepts behind Brunt's method, its assumptions, and an implementation of it. A detailed derivation of the method and the quantities mentioned here can be found in Appendix C.

The main idea behind the method is to reconstruct the properties of a three-dimensional source from the information contained in its observed two-dimensional line-of-sight projection. Assuming that the observed source is described by the three-dimensional field $\mathbf{F}$, its two-dimensional projection (average along one axis, the z-axis in this case) is denoted by $\mathbf{F}_p$. In Appendix C it is shown that the Fourier transform $\tilde{\mathbf{F}}_p$ of $\mathbf{F}_p$ is proportional to the $k_z = 0$ cut of the transform $\tilde{\mathbf{F}}$ of $\mathbf{F}$,

$$\tilde{\mathbf{F}}_p(k_x, k_y) = \tilde{\mathbf{F}}(k_x, k_y, k_z = 0). \tag{5.1}$$

If $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{F}}_p$ only depend on the wavenumber $k = |\mathbf{k}|$ (isotropic fields), the average properties of $\mathbf{F}$ can be derived from their two-dimensional counterparts of $\mathbf{F}_p$ through symmetry arguments. When a field such as the velocity or the momentum is measured in observations, only its line-of-sight component is available A two-dimensional projected field is recovered by considering the Helmholtz decomposition of the line-of-sight component. According to the Helmholtz theorem, a vector field can be split into a divergence-free (solenoidal or transverse) component, $\mathbf{F}_\perp$ and curl-free (compressive or parallel) component, $\mathbf{F}_\parallel$. In Fourier space, the solenoidal and compressive components are linked through (local) orthogonality. As the name suggests the divergence-free

(solenoidal) component encodes the turbulent, vorticose modes of a flow. Compressive modes, accounting for compression and expansion of the gas are embodied by the curl-free component. These modes are likely to be connected to star-formation. To obtain a unique decomposition, the vector field must satisfy suitable boundary conditions (the Helmholtz decomposition is defined up to a vector constant, see Appendix C). In particular, it is required that the field decay to zero smoothly on the boundary. This condition also ensures that the Fourier transforms of the observed field are well-behaved as these fields are not naturally periodic. Isolated, gravitationally bounded molecular clouds possess a natural boundary, however, when the signal is truncated artificially by the edges of the observed field, apodisation of the emission at the edge is required to restore a suitable boundary. As mentioned above, statistical isotropy is also required for the method to be applied. Sources of strong anisotropy such as strong magnetic fields or filamentary shapes thus heavily affect the reliability of the results. Fields with steep power spectra should also be avoided. In practice, such power spectra show high sensitivity to low spatial frequencies which are poorly sampled statistically.

Assuming the emission line under consideration is optically thin and that the emissivity depends solely on the volume density, the PPV datacube can be translated into a density weighted field spanning the region of observation. This field is the 'momentum density' (see Appendix C)

$$\mathbf{p} = \rho \mathbf{v}, \tag{5.2}$$

composed of the volume density ρ and the velocity field $\mathbf{v}$.

The ratio of the variance of transverse momentum density to the variance of the total momentum density gives the solenoidal fraction, R. This fraction represents the amount of power in the solenoidal modes of the momentum density in a given region of space,

$$R = \frac{\sigma^2_{p_\perp}}{\sigma^2_p}. \tag{5.3}$$

Brunt & Federrath (2014) demonstrated that the solenoidal fraction can be expressed in terms of observable quantities: the zeroth, first, second velocity moments, and their power spectra. The first three velocity moments are defined as

$$W_0 = \int I(v)\ dv, \qquad W_1 = \int vI(v)\ dv, \qquad W_2 = \int v^2 I(v)\ dv. \tag{5.4}$$

With the assumption that the thermal linewidth is negligible compared to the overall velocity dispersion, the velocity moments can be recast in terms of density (Brunt & Federrath, 2014)

$$W_0 \propto \int \rho(z)\ dz, \qquad W_1 \propto \int v(z)\rho(z)\ dz, \qquad W_2 \propto \int v(z)^2\rho(z)\ dz. \tag{5.5}$$

These moments allow for the solenoidal fraction to be written as

$$R = \left[\frac{\langle W_1^2\rangle}{\langle W_0^2\rangle}\right]\left[\frac{\langle W_0^2/\langle W_0\rangle^2\rangle}{1 + A(\langle W_0^2\rangle/\langle W_0\rangle^2 - 1)}\right]\left[g_{21}\frac{\langle W_2\rangle}{\langle W_0\rangle}\right]^{-1} B, \tag{5.6}$$

where

$$A = \frac{(\sum_{k_x}\sum_{k_y}\sum_{k_z} f(k)) - f(0)}{\sum_{k_x}\sum_{k_y} f(k)) - f(0)}, \tag{5.7}$$

and

$$B = \frac{(\sum_{k_x}\sum_{k_y}\sum_{k_z} f_\perp(k)\frac{k_x^2+k_y^2}{k^2})}{\sum_{k_x}\sum_{k_y} f_\perp(k)}, \tag{5.8}$$

with $f(k)$ and $f_\perp(k)$ being the angular (azimuthal) averages of the power spectra of the zeroth and first moments (notation after Orkisz et al., 2017). The constant g_{21} is a statistical correction factor that accounts for the correlations between the variations of ρ and $\mathbf{v}$ (if ρ and $\mathbf{v}$ are not correlated, $g_{21} = 1$). In terms of density, velocity and the spatial average of the density ρ_0, g_{21} is expressed by the variance of the three-dimensional volume density $\langle(\rho/\rho_0)^2\rangle$ as

$$g_{21} = \frac{\langle\rho^2 v^2\rangle/\langle\rho^2\rangle}{\langle\rho v^2\rangle/\langle\rho\rangle} = \left\langle\frac{\rho^2}{\rho_0^2}\right\rangle^{\epsilon}. \tag{5.9}$$

The exponent ϵ is a is a small positive constant which is the exponent of the power law expressing the relation between the variance of the velocity σ_v^2 and the density ρ (see section 5.2.6).

In the hypersonic regime ($M > 5$) the solenoidal fraction becomes independent of the type of forcing and converges to $R \sim 2/3$ (Brunt & Federrath, 2014). This specific value reflects the equipartition of momentum between the compressive and solenoidal mode Federrath et al. (2008a). Values of the solenoidal fraction that are higher than 2/3 imply that the relative fraction of momentum density in solenoidal modes in the flow exceeds that in compressive modes. Thus, star formation tends to be suppressed.

5.2 Application

5.2.1 Observations

The method described above is applied to a selection of SCIMES clouds extracted from the CHIMPS ^{13}CO (3-2) emission data. The reduced catalogue is constructed through a size criterion that selects sources with a spatial extension of at least 9 voxels in each direction and a spectral width of at least 1 $\mathrm{km\,s^{-1}}$. This choice allows for a minimum resolution of 4 times the size of the telescope beam. This constraint ensures the inclusion of sources that extend well above the telescope resolution and exclude possible artefacts and very narrow filamentary structures. In addition, the smallest clouds in this selection are large enough to include an envelope of rarefied gas around the densest, brightest peaks. This supports our assumption of considering ^{13}CO ($J = 3 \rightarrow 2$) to be optically thin in diffuse regions (with optical depth increasing around the peaks of emission, where the cloud is densest, Rigby et al., 2016). In a typical cloud, the volume occupied by the diffuse component far exceeds the denser parts.

The selected sub-catalogue includes a few very large clouds, two of which contains tens of millions of voxels. In these cases, the calculation of the power spectra of the velocity moments becomes cumbersome and resource-demanding. To avoid impractically long computation times, deterioration the resolution is applied to such clouds by a factor of 2 on each axis.

5.2.2 Isolating the clouds

The emission of each cloud in the selection is isolated via a mask constructed from the SCIMES clusters assignment catalogue. The resulting map is used for the computation of moments (see Figure 5.1).

5.2.3 Moments

The velocity moments whose power spectra enter the formula for the solenoidal fraction must be calculated in the frame of reference of the centre of mass of the cloud. Thus, to express the velocity moments in the centre-of-mass frame, first, the centroid velocity of the cloud in the LSR frame is calculated. This quantity is simply given by the ratio

$$V_c = \frac{\langle W_1^{\rm obs} \rangle}{\langle W_0 \rangle}, \tag{5.10}$$

of the spatial means of the first moment in the observer's frame and $\langle W_1^{\rm obs} \rangle$ and of the zeroth moment $\langle W_0 \rangle$. Notice that, not being velocity weighted W_0 is invariant of the frame of reference. The resulting change of coordinates gives

$$v = v_{\rm obs} - V_c \tag{5.11}$$

(adopting the same notation as before). Finally, substituting in the first and second moments yields

$$W_1 = \int (v_{\rm obs} - V_c) I(v_{\rm obs})\ dv_{\rm obs} \tag{5.12}$$

and

$$W_2 = \int (v_{\rm obs} - V_c)^2 I(v_{\rm obs})\ dv_{\rm obs}. \tag{5.13}$$

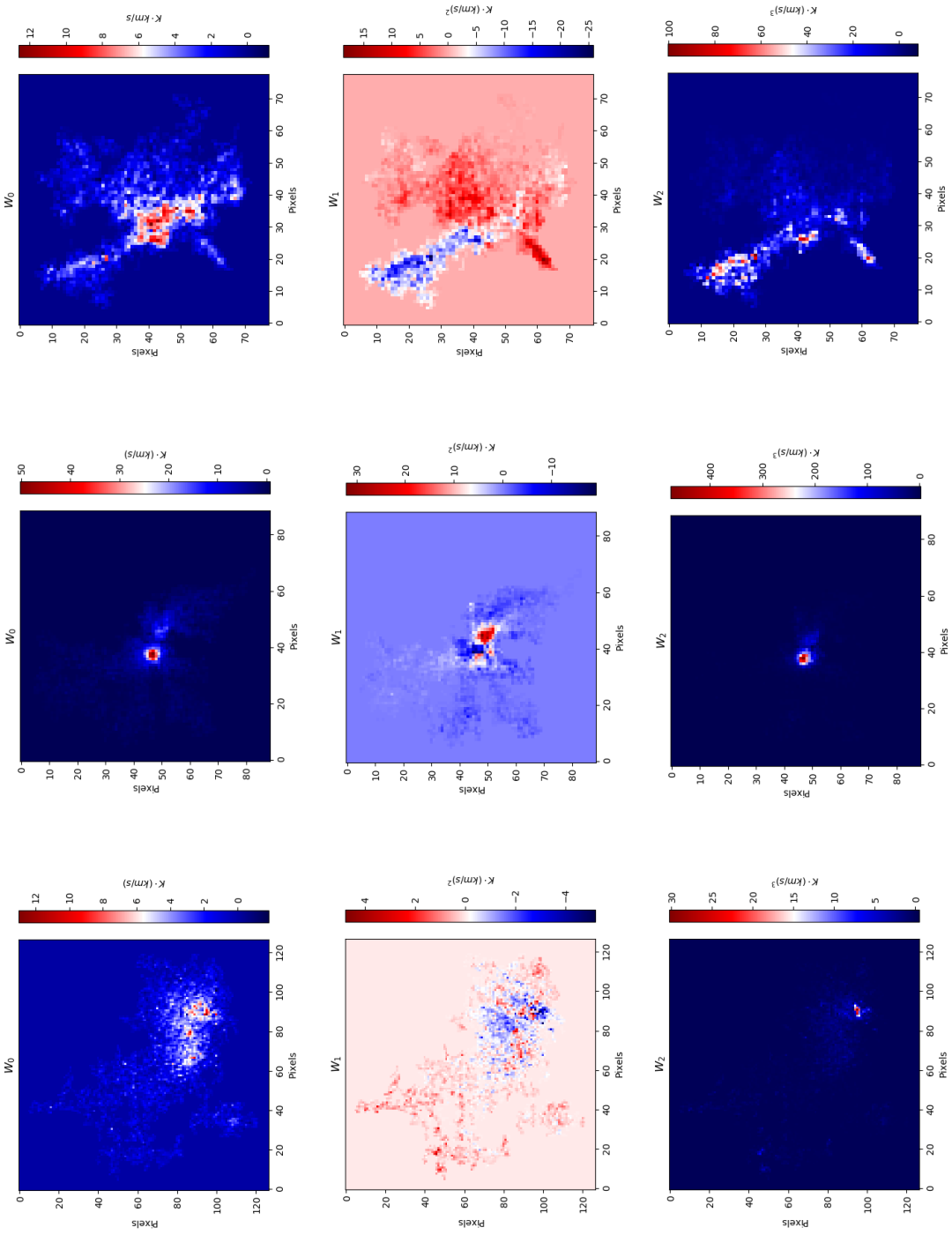


FIGURE 5.1: Moment maps of three molecular clouds. Each column includes the moment maps of single cloud. The panels W_0, W_1, and W_2 depict the integrated intensity along the spectral axis (zeroth moment), the first and second moments of the velocity respectively.

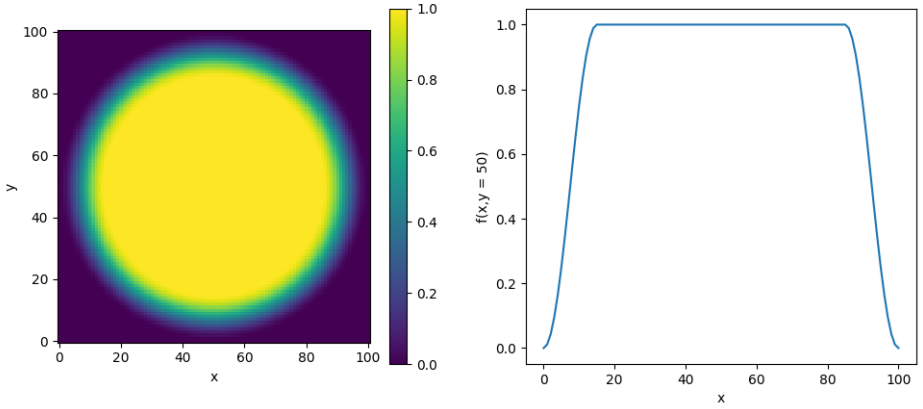


FIGURE 5.2: The Tukey window over 101×101 square. The Tukey window is a rectangular window which equals a cosine function over the first and last $r/2$ percent of its domain[1]. The cosine fraction r regulates the shape of the Tukey window. For instance, a Tukey window with $r = 0.5$ has segments of a phase-shifted cosine with period $2r = 1$ that cover half of the length of the window. The figure was generated with the code provided in TurbuStat documentation[2].

5.2.4 Padding and apodisation

Once the moment maps of a cloud have been constructed, the cloud is extracted by enclosing it into a square region of the map. The size (side) of this region is determined by considering the maximum extension of the cloud along the coordinate axes with an added 5-pixel padding in every direction. This ensures that the moment field is zero at the edges of the region. For clouds that touch the edges of the field of observation, an artificial boundary is created. In this case, apodisation is required to ensure that field decays to zero at the edges. A Tukey window with a cosine fraction equal to 0.3 is used as apodising kernel (see Figure 5.2). This kernel was found to be the most efficient at smoothing out high-frequency artefacts in the clouds considered. However, applying an apodising kernel affects the power spectrum of an image (see the section below). The range of frequencies affected by the kernel depends both on the properties of the kernel used and the features of the map. Narrower shapes usually have a bigger impact on the power spectrum.

Apodisation with a Tukey window may bias the shape of the power spectrum at large frequencies, usually over scales above 1/2 of the map size (Koch et al., 2019).

5.2.5 Power spectra

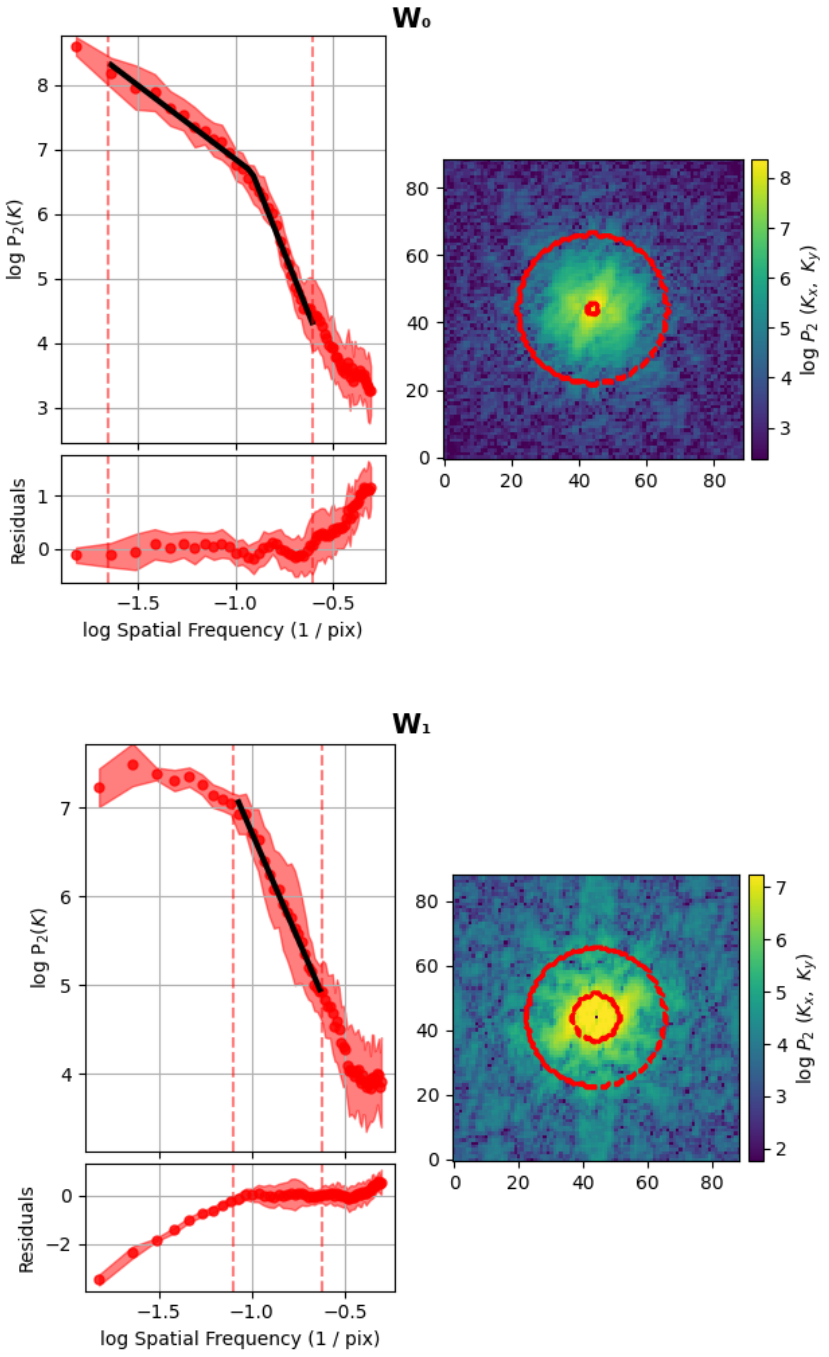
The power spectra of the moments maps are calculated by the PowerSpectrum method of Turbustat (Koch et al., 2019). Turbustat is a Python package that implements a suite of tools devoted to the statistical analysis of turbulence[3]. PowerSpectrum implements a model for the computation of the full two-dimensional spatial power spectrum of an image (elliptical power-law model). A radial profile of the two-dimensional power spectrum produces the azimuthally averaged one-dimensional power spectrum that is required for the calculation of the solenoidal fraction. PowerSpectrum also provides a power-law fit for a one-dimensional power spectrum. Different physical processes characterising distinct scales may induce breaks in the power-law behaviour of the power spectrum. PowerSpectrum accounts for this situation through fitting with a segmented linear model (Figure 5.3). An initial guess of the scale of the breaking point can be passed to the power spectrum. The segmented linear model then attempts to optimise the frequency of the breaking point by minimising the gap between the two individual linear components. If no good location for the breaking point is found, PowerSpectrum adopts a linear fit for the entire spectrum. An optimised breaking point parameter is useful to understand the scales of different regimes in the turbulence which are characterised by specific slopes [4].

To avoid large deviations on small scales (high spatial frequencies) where the information has been lost by the spatial smoothing applied to the image (convolution of the beam), only spatial frequencies that correspond to twice the FWHM value of the telescope beam are considered.

This correction also mitigates the impact of the noise which is more severe at higher spatial frequencies. Modelling the power spectra of the observable moments as the sum of the beam-convoluted signal spectrum and a noise spectrum (Brunt et al., 2010; Orkisz et al., 2017), the amplitude of the noise component is expected to be several orders of magnitude smaller than the signal spectrum, becoming comparable in magnitude at

[3] https://turbustat.readthedocs.io/en/latest/index.html

[4] Kolmogorov turbulence, for instance, obeys a power law with exponent $k = -5/3$, while $k = -2$ characterises Burgers' turbulence. As observations depend both on velocity and density, the exponent of the power spectrum of an integrated intensity map will also depend on the optical depth of the gas (and the fluctuations in both fields) (Lazarian & Pogosyan, 2000)). Optically thin and optically thick gas saturates at $k = -3$ and $k = -11/3$ respectively (Lazarian & Pogosyan, 2004; Burkhart et al., 2013).



centring

FIGURE 5.3: The power spectra of the zeroth (W_0) and first (W_1) moment maps. Each panel shows both the angular averaged 1D and full 2D power spectra. The dashed lines in the one-dimensional spectra and the corresponding red circles in the two-dimensional power spectra delimit the region over which the spectrum is fitted with a segmented linear model. The fitted power-law model of the 1D spectrum is denoted by the solid black line.

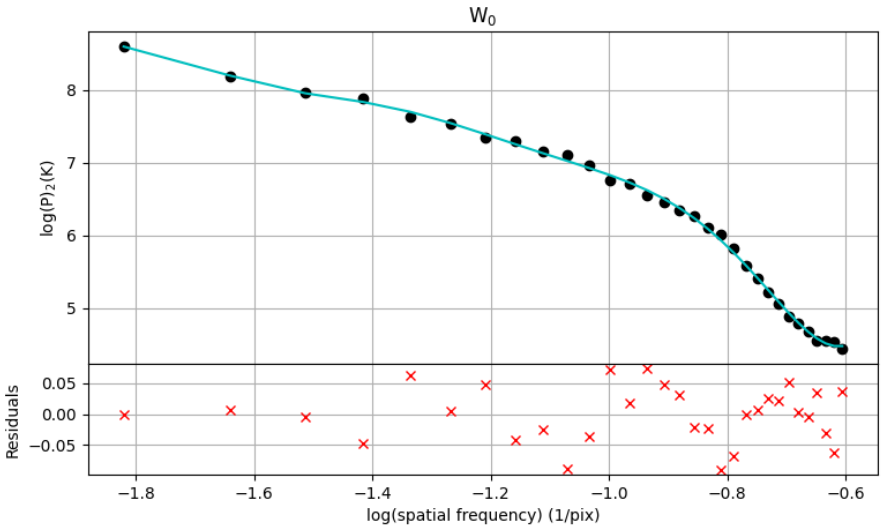


FIGURE 5.4: Polynomial fit (7th grade) of the full azimuthally averaged power spectrum of the zeroth moment.

frequencies around the telescope resolution [5]. Deconvolution by the beam is performed automatically by PowerSpectrum, but no corrections for padding (Brunt et al., 2010) were used in the analysis in Chapter 6. In addition, over-sampling of the beam generates an increase in power at high frequencies. This region should also be omitted from the fit of the power spectrum. Thus power laws alone are not sufficient to obtain an accurate fit over the entire spectrum. To approach this problem a tentative interval (and a breaking point) over which to apply the segmented linear model is identified. This is accomplished by fitting the entire power spectrum with a seventh-degree polynomial and studying its local extremals to isolate a region of descending slope. The breaking point is chosen as the mid-point of this interval. PowerSpectrum is then re-run on the cut data that cover this interval. The power spectrum on frequencies outside the interval is fitted by linear interpolation or polynomial fit (Figures 5.4 and 5.5).

To recover all values of the wave vector components that appear in the summations in equation 5.6 from the wave vector bins of the azimuthally averaged power spectrum, a fitting algorithm for the one-dimensional power spectrum was devised. Linear interpolation of the data points obtained through PowerSpectrum is employed as the best approximation of the power spectrum. Applying different fitting functions affects the resulting value of the solenoidal fraction. The goodness of fit determines the size of the

[5] This behaviour appears in several SCIMES clouds with noise spectra estimated in survey areas where emission is absent. However, for the time being, the hypothesis has not been validated for the entirety of the CHIMPS sample.

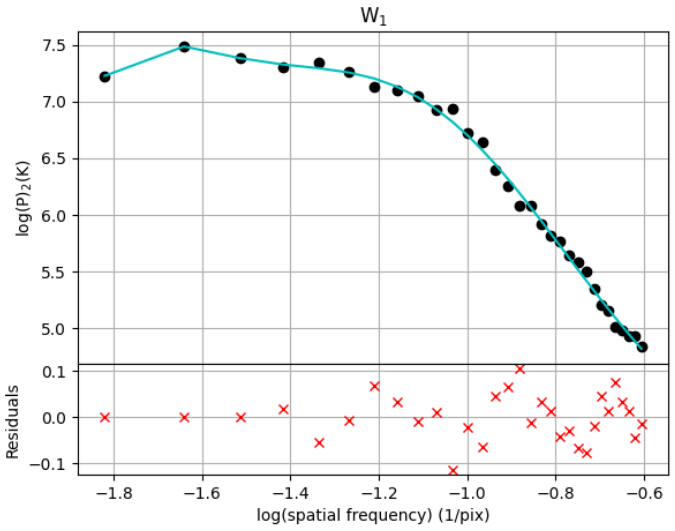


FIGURE 5.5: Polynomial fit (7th grade) of the full azimuthally averaged power spectrum of the first moment.

variation in the solenoidal fraction. Comparing full interpolation to a 7th-degree polynomial fit, an average difference of $\sim 4\%$ in the solenoidal fraction is found. This value rises to 13% using a combination of linear fitting (power-law estimated by PowerSpectrum) and interpolation outside the linear regime (power law) region. This choice addresses the method's sensitivity issues at large spatial scales (low frequencies of the power spectrum) due the characteristics of the sums in the parameters A and B (equations C.98, C.99) reported in Brunt & Federrath (2014).

The 2D dimensional power spectra of the zeroth and first moments do not show any marked anisotropy. Assuming that power spectra are isotropic in the spatial dimensions, then statistically the third dimension is expected to follow this isotropy as well. Therefore, fulfilling the isotropic requirement.

5.2.6 Density-velocity correlations

The exponent ϵ in equation 5.9 is set to 0.15 This value was derived by Orkisz et al. (2017) for their analysis of the solenoidal fraction in Orion B. Their estimation of the relation linking local density and velocity dispersion is based on several emission lines with different spatial distributions in the mean spectrum (mean line profiles). They considered five isotopologues to trace gas at different densities:

- ^{12}CO($J = 1 \rightarrow 0$) and HCO$^+$($J = 1 \rightarrow 0$) for low density gas (Pety et al., 2017),
- ^{13}CO($J = 1 \rightarrow 0$) for the bulk of the cloud (Orkisz et al., 2017),
- C^{18}O($J = 1 \rightarrow 0$) for denser and shielded regions (Hily-Blant et al., 2005),
- N_2H^+($J = 1 \rightarrow 0$) for the densest cores (Kirk et al., 2016).

These lines may all appear in the emission from gas at different densities. However, there is a density lower bound past which a given transition vanishes. Below this density threshold, the molecule may either not be present or not be excited. A density threshold that corresponds to the velocity dispersion of the emission line is taken. The velocity dispersion (FWHM) of lines of these species was determined by fitting of a Gaussian line profile or using the information on the hyperfine structure of the molecule (N_2H^+) [6].

Orkisz et al. (2017) devised an empirical relation between the fitted velocity dispersion velocities (δv) and lowest emission density ($\rho(H_2)$) from the data of the five species:

$$\delta v \propto \rho(H_2)^{-0.15}. \tag{5.14}$$

The slope $\alpha = -\epsilon = -0.15$ is derived from a least-squares fit of the variation of the FWHM with the density. Orkisz et al. (2017) estimated that possible systematic errors in the ^{12}CO (1-0), HCO$^+$ (1-0), and N_2H^+ (1-0) densities and the ^{12}CO (1-0) and HCO$^+$ (1-0) linewidths tend to steepen the slope of the power law. Thus, $\epsilon = 0.15$ should be considered as an upper bound. This value corresponds to an upper bound of the correction factor g_{21} (equation 5.9). A lower bound of g_{21} is provided by $\epsilon = 0.05$ as estimated by Brunt & Federrath (2014).

5.3 Summary

This chapter provides a recipe for the calculation of the solenoidal fraction in molecular emission datasets. It includes a brief overview of the method (Brunt et al., 2010; Brunt & Federrath, 2014) and an introduction of the equations and their terms (that are fully

[6] `http://www.iram.fr/IRAMFR/GILDAS`.

derived in Appendix C). Finally, the recipe that constitutes the core algorithm for the calculation of the solenoidal fraction is provided.

Chapter 6

Analysis of turbulence: Results

The nature of turbulence, and distinctly its solenoidal or compressive modes, is hypothesised to be a factor in the collapse of dense gas regions in molecular clouds (Federrath & Klessen, 2012), thus playing a part in the star-formation efficiency of individual clouds which is observed to vary by 2-3 orders of magnitude (Eden et al., 2012, 2013; Rigby et al., 2016). In particular, compressive flows are linked to typical aspects of star formation such as gas infall on filaments, the collapse of dense cores, and the expansion around young stars. Thus, a cloud dominated by compressive turbulence can be expected to be more likely to host collapsing regions and consequently have a higher star formation efficiency (Federrath & Klessen, 2012). A study of the Orion B molecular cloud (Orkisz et al., 2017) finds that the overall turbulent modes are mostly solenoidal, consistent with the observed low star formation rate. However, the turbulent modes estimated are position-dependent and vary with scale within the cloud, with motions around the main star-forming regions being strongly compressive. Although this analysis confirms that a high solenoidal fraction (see 5.1) means a dominant non-compressive forcing and suggests that star formation is less efficient in the case of the Orion B complex, a full sample study of the relation between turbulent modes and star formation efficiency is still missing.

The method introduced in Chapter 5 (and described in full in Appendix C) is here

applied to a selection of CHIMPS clouds identified with the SCIMES method (Chapters B.5 and 4). The sample under investigation is selected through size constraints (section 5.2.1) and includes 1311 isolated clouds, 963 of which are associated with an independently measured star formation efficiency.

6.1 The solenoidal fraction

The sample selection described in section 5.2.1 produces a collection of 1311 SCIMES clouds, of which 1283 are isolated clouds, while 28 cross the edges of the field of observation. This latter set of clouds require apodising (see 5.2.4). Although most of these clouds are small and located across the latitude boundaries, some of them have significant sizes, covering fairly large areas between CHIMPS regions (especially regions 1 and 2 at longitudes between 30°.5 and 32°).

The solenoidal fraction (introduced in Chapter 5 and Appendix C), is calculated through an algorithm that automates the steps described in sections 5.2.2, 5.2.3, 5.2.4, 5.2.5, 5.2.6 allowing for the method to applied to a large sample. This algorithm produces the value of the solenoidal fraction associated with each cloud in a SCIMES cluster assignment map, given its corresponding cloud catalogue, the survey emission, and column density data as input. Apodisation is performed for those clouds crossing the edges of the field of observation (EDGE labels). A polynomial fit of the power spectrum is run to determine the domain of the power law fit, which is then carried out with Turbustat while the ends are fitted by interpolation.

Figure 6.1 shows the distributions of solenoidal fraction for the sub-samples with and without associated Hi-GAL bolometric luminosities (see section 6.2). These distributions appear to show that the sample without associated luminosities is shifted to slightly higher solenoidal fractions. This behaviour is consistent with the hypothesis that a higher solenoidal fraction reduces the likelihood of star formation. To check if the sub-samples are significantly different a Kolmogorov-Smirnov test is performed over the two distributions of the solenoidal fraction. Following the convention set in kstest in the package Scypy, with the null hypothesis that the two samples (distributions) are drawn from the same distribution, while the alternative is that they are independent. The test returns $k = 0.44$ with p-value $= 2.11 \times 10^{-15}$, the null hypothesis can thus

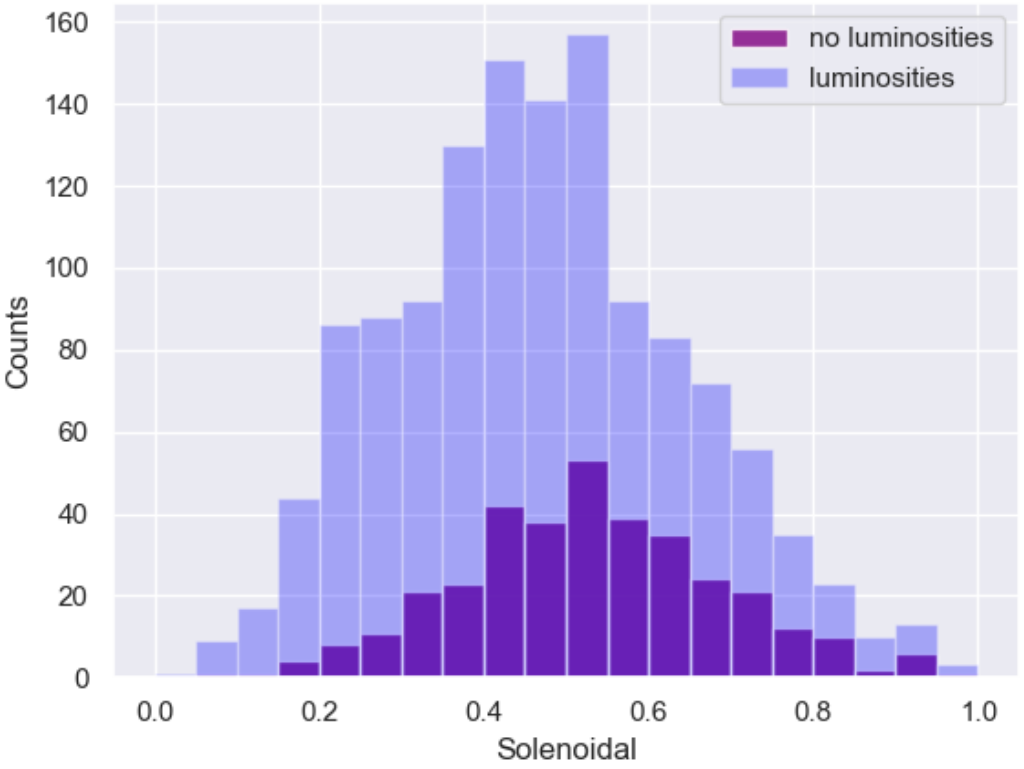


FIGURE 6.1: Distribution of solenoidal fraction within the size constrained sample of 1311 SCIMES clouds (blue histogram). The purple histogram traces the distribution of the subset of sources that do not have Hi-GAL luminosity counterparts (see section 6.2).

be rejected. The lack of Hi-GAL luminosity for 350 sources depends on the missing detection in the Hi-GAL IR bands (in the full merged catalogue, see section 2.4). In particular, the lack of 70 μm emission is commonly considered a sign of no embedded star formation (or at least, star formation that is not detected). Inaccuracies in the assignment of Hi-GAL luminosities to SCIMES sources (see Section 6.2) may also affect the distribution shown in Figure 6.1. Hi-GAL sources lack velocity measurements so that the luminosity assignment must be performed through line-of-sight projections which may cause blending of near-far luminosities. In addition, the coordinates of Hi-GAL are given with respect to the emission features identified by the CUTEX algorithm (see section 2.4. The discrepancies between the CUTEX and SCIMES extractions will also result in the loss of precision in the luminosity assignments.)

An error estimation in the solenoidal fraction was performed by comparison between the

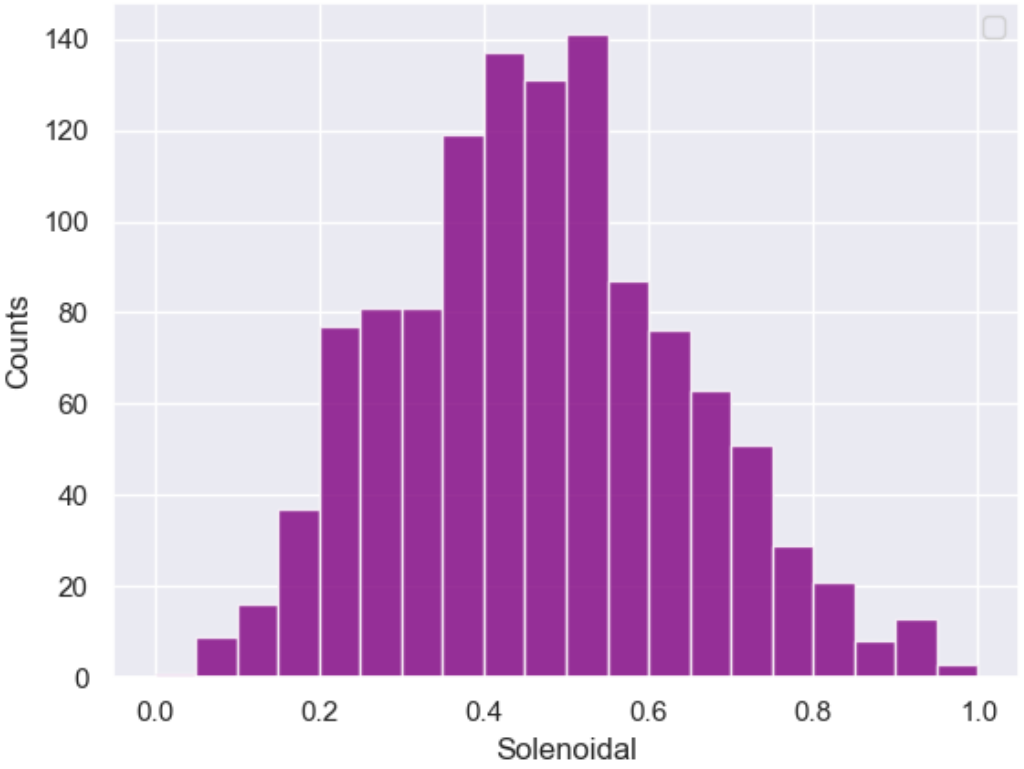


FIGURE 6.2: Distribution of solenoidal fraction for clouds in hyper-sonic regimes (Mach number > 5). This sub-sample comprises the 92% (1218 sources) of the original selection for which the solenoidal fraction is calculated. With solenoidal fractions $< 2/3$ the majority of hypersonic clouds have the potential to form stars.

original catalogue and the calculation on emission maps perturbed by the addition of the square root of the corresponding variance maps. The method returned an average error of 7% which is consistent with the 8-13% ranges found in the Orion B emission (Orkisz et al., 2017).

Brunt & Federrath (2014) showed that, theoretically, at hypersonic regimes (Mach numbers ~ 5) the solenoidal fraction of the momentum density becomes independent of the type of forcing and converges to 2/3. This value follows from the equipartition of momentum between the solenoidal and compressive modes (Federrath et al., 2008a). A solenoidal fraction smaller than 2/3 implies a loss of equilibrium in favour of the compressive modes of the flow. When this situation occurs, a cloud is more likely to form stars.

Isolating the subset of sources in hypersonic regimes reveals (see Figure 6.2) that this

selection comprises 92% of CHIMPS sources for which the solenoidal fraction has been calculated. In turn, only 4% of hypersonic sources have $R > 2/3$, thus most of the original selection has the potential to form stars. Values of the solenoidal fraction that exceed 2/3, in this case, may be caused by systematics and measurement errors. These fractions imply that the result is free of potential concerns over the nature of the forcing mechanism being a factor in the value of the solenoidal fraction. At these sonic regimes, complete mixing of turbulent modes is expected. Values lower than the 2/3 ratio can either indicate a specific forcing for the turbulent flow at low Mach numbers (transonic regime, $0.8 < M < 1.2$), or suggest that an ordered flow is superimposed on the mixed turbulence at high Mach number (Brunt & Federrath, 2014). Only a small fraction of clouds have transonic velocities, so the forcing mechanism does not appear to be a factor in determining the solenoidal fraction for this sample. It follows that the solenoidal fraction is more likely to be set by the superimposed ordered flow (collapse or outflow resulting from star formation).

Figure 6.3 shows the distribution of solenoidal fraction with Galactocentric distance. The width of the bins is 0.5 kpc until 8 kpc and 1 kpc from 8 to 10 kpc and 2 kpc past this distance. The reason for using irregular bin widths is to reduce biases by considering bin populations of similar sizes. Bin widths are represented by the length of the horizontal blue lines that indicate the mean value of the solenoidal fraction in each bin.

The solenoidal fraction peaks at the $3 - 4$ kpc bin. If confirmed by the analysis of a sample at lower longitudes, this result would be consistent with the disc becoming stable against gravitational collapse. This distance marks the boundary of the inner Galaxy, the region of influence of the Galactic bar, which in extragalactic systems has been observed to quench star formation (see section 7.2 in the next Chapter).

The number of clouds with distances smaller than 4 kpc amounts to 8. These clouds have projected sizes ranging from 81 to 1640 pixels (with an average of 508) and field sizes from 33 to 108 pixels and averaging at 63 (and including two clouds with field sizes above 85 pixels, see section 6.4). This set of clouds does not present any special, unique size-related features and is consistent with the entire population. Visual inspection of their size agrees with the Kolmogorov-Smirnov test ($k = 0.18$ and p-value $= 0.92$) proving that these clouds were sampled from the full distribution. The small size of

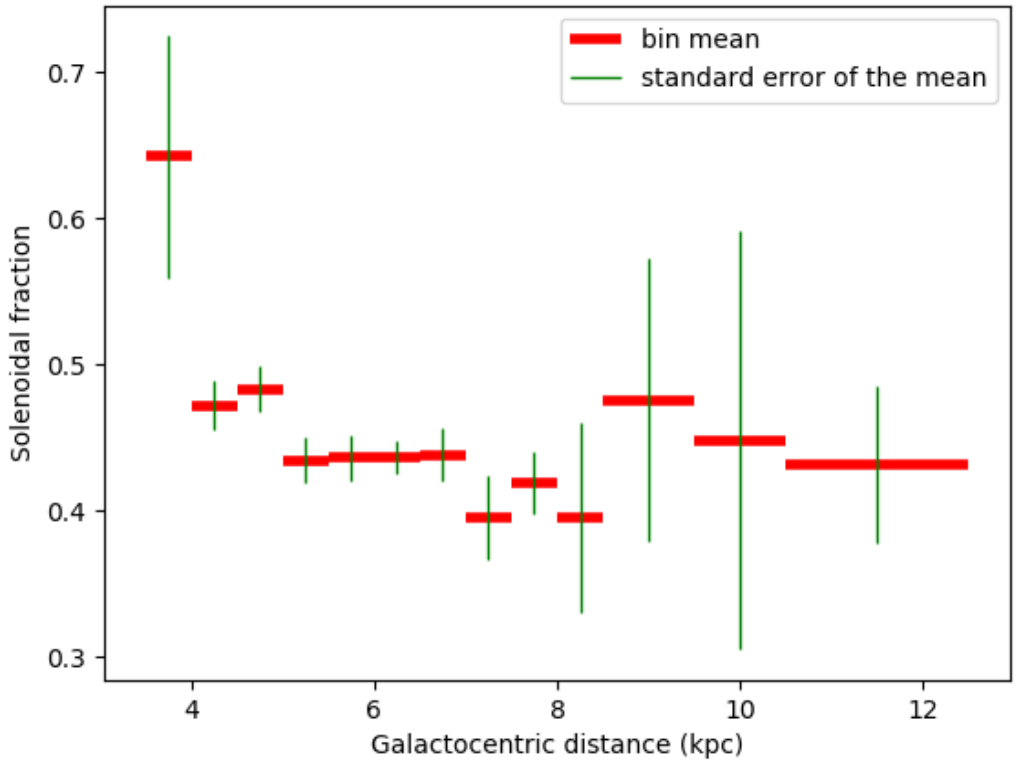


FIGURE 6.3: Distribution of the solenoidal fraction with the Galactocentric distance. The size of the bins is adjusted to the number of sources. The bins are 0.5 kpc wide until 8 kpc and 1 kpc wide from 8 to 10 kpc. At distances larger than 10 kpc, clouds are collected in a single 2 kpc bin. The horizontal blue lines indicate the mean value within each the bins. The vertical bars represent the standard error of the mean.

the set makes this point of low significance but nonetheless invites further work at low longitudes.

The solenoidal fraction then declines with a shallow gradient with increasing Galactocentric distance. For Galactocentric distances greater than 4 kpc, a Spearman test returns $r = -0.133$ with p-value $= 1.498 \times 10^{-6}$ indicating that the solenoidal fraction declines with distance from the Galactic centre. This decrease corresponds to a shallow gradient with a slope of -0.02 with no signal present at the spiral-arm radii. This result is in agreement with previous studies that found no significant arm associated signal (Ragan et al., 2016, 2018). Figure 6.4 shows the distribution of the solenoidal fraction with heliocentric distance.

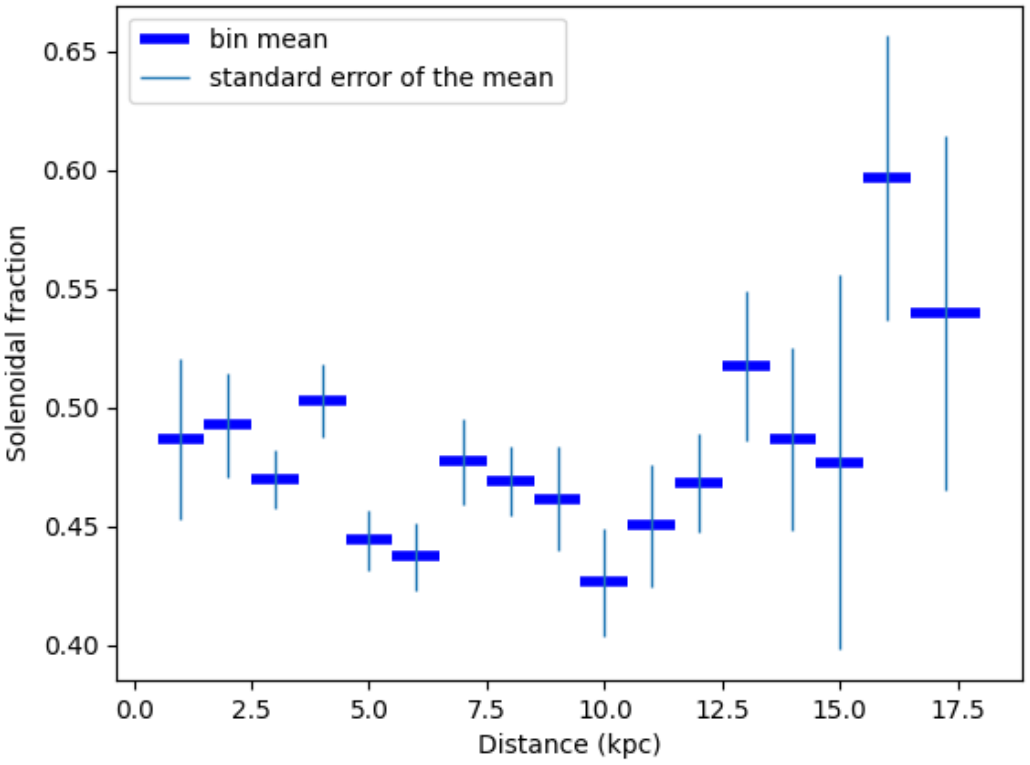


FIGURE 6.4: Distribution of the solenoidal fraction with the heliocentric distance. The size of the bins is adjusted to the number of sources. The bins are 1 kpc wide until 16 kpc and 1 kpc wide thereafter. The horizontal blue lines indicate the mean value within each the bins. The vertical bars represent the standard error of the mean.

No significant correlation (Spearman statistics) was found between the solenoidal fraction, mass, and Mach numbers. In particular, the solenoidal fraction is not correlated to the volume of the clouds (number of voxels) ensuring that the results are not affected by resolution biases.

These results suggest that the state of the physical properties of a cloud and thus its likelihood to form collapsing cores may be linked to the Galactic environment or individual cloud formation histories in which the cloud is located, slowly changing in the disc and possibly steepening into the bar-swept region and continuing into the CMZ which has very low SFE (Longmore et al., 2013; Urquhart et al., 2013).

6.2 Star formation efficiency

Star formation efficiency (SFE) can be understood as the fraction of dense, ^{13}CO (3-2)-traced clouds/clumps that have collapsed and turned into stars over some time-scale (Eden et al., 2015). The YSO luminosity as a function of time may represent the star-formation history of a cloud. With this notion of efficiency, SFE can then be defined as the ratio of the IR luminosity of the YSOs embedded in a cloud to the mass of the cloud:

$$\mathrm{SFE} = \frac{L_{\mathrm{star}}}{M_{\mathrm{cloud}}} = \frac{1}{M_{\mathrm{cloud}}} \int_0^t \frac{dL}{dt} dt, \tag{6.1}$$

where dL/dt is the instantaneous star formation rate (SFR) in terms of the integrated luminosity L of YSOs. Large values for L/M are either due to a high SFR or a long time scale. Therefore to directly identify L/M with the SFE requires the assumption that dL/dt be proportional to dM/dt (linear dependence), which in turn necessitates that the stellar IMF is invariant and fully sampled in all star-forming regions, up to the maximum stellar masses (Weidner & Kroupa, 2006). If, more realistically, the IMF is filled stochastically (Elmegreen, 2006), then the L/M may depend on SFE non-linearly. In this case, an increase in the SFE still corresponds to an increase in L/M. An observed rise in L/M may however also be produced by the formation of a larger star cluster with a more fully sampled IMF in larger clouds. For clusters, L is proportional to M^2 with the same SFE. This potential variation in L/M cannot be resolved by observations unless it is possible to distinguish every single star in the cluster (which is beyond the limits of current technology). When the SFE is high, non-linearity may be caused by variations in the mass of the cloud. As SFE is generally lower than 30% (Lada & Lada, 2003), M can be assumed to remain constant over the time-scale of star-forming events detected in the mid-and far-IR. In theory, L/M evolves with time too (increasing L and decreasing M) and it becomes necessary to define the SFE in terms of a specific time-scale (e.g. free-fall time, see Cheavance, 2020). On the other hand, the stage of massive star formation that can be detected in the mid-IR and far IR lasts for only hundreds of thousands of years (Davies et al., 2011; Mottram et al., 2011), a short enough time to allow us to consider L/M can be as a "snapshot" of the current SFE.

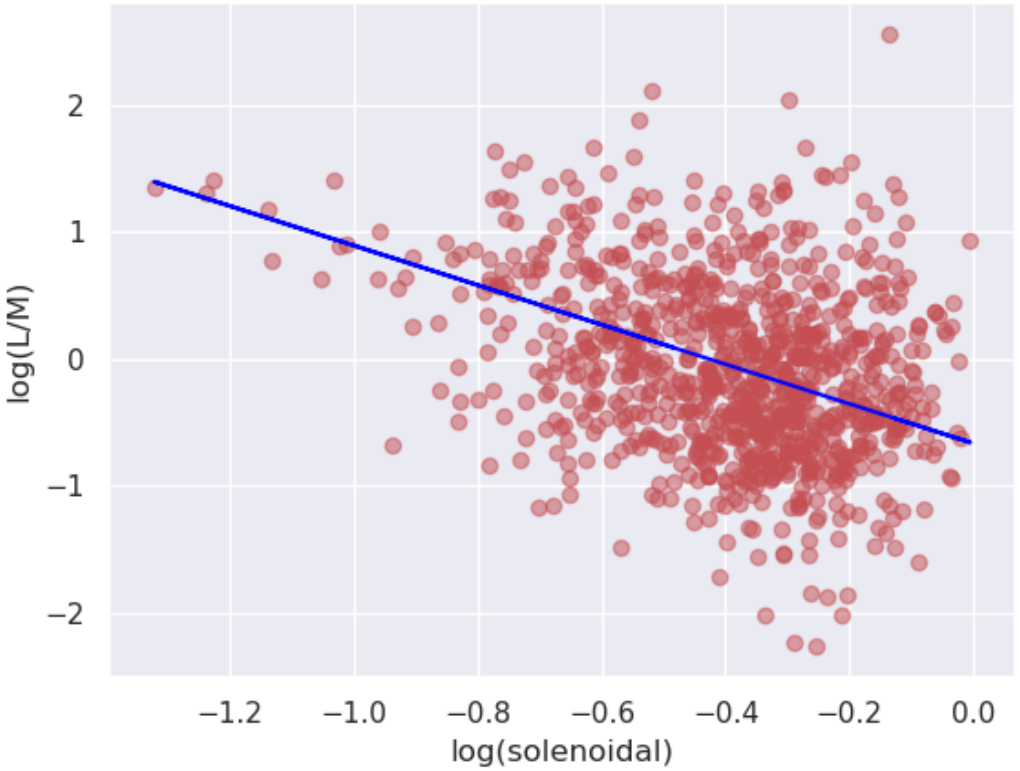


FIGURE 6.5: Star formation efficiency defined as L/M (in units of Solar mass and Solar luminosity) as a function of the solenoidal fraction. The IR continuum luminosity from the YSOs and the masses of the sources (CO mass) are derived from independent measurements and can be considered largely independent variables. The cyan solid line is a weighted linear fit of the scatter plot. The weights are the standard deviations of the L/M distribution within solenoidal fraction bins with width 0.1.

Luminosity assignments are made using the integrated bolometric fluxes of the Hi-GAL sources contained within each SCIMES cloud. Since the Hi-GAL catalogue does not include velocity information, a Hi-GAL source is matched to a SCIMES cloud when its Galactic coordinates lie within the projection of the SCIMES cloud on the Galactic plane. This assignment however is not (always) unique as projecting along the spectral direction may result in the full or partial overlapping of multiple SCIMES clouds. The position of a Hi-GAL source on the Galactic plane may thus belong to several distinct projected clouds. When this happens, the assignment is made unique by associating a Hi-GAL source with the SCIMES cloud that has the brightest ^{13}CO (3-2) intensity along the spectral direction at the source's coordinates. This method allows us to define a luminosity for 963 clouds in the original sample.

A negative correlation (Spearman $r = -0.30$, p-value $4.23\dot{1}0^{-21}$) is found between star formation efficiency and the solenoidal fraction. The correlation in Figure 6.5 is again consistent with the hypothesis that star formation is more likely to occur in clouds with more power in the/more dominant solenoidal turbulent modes. The IR continuum luminosity from the YSO and the CO mass of the cloud were derived from independent measurements. They can therefore be treated as largely independent variables, making the correlation valid as potentially revealing physical effects. This is however not the case when M itself is also based on the continuum emission (Molinari et al., 2016; Urquhart et al., 2018). With this derivation of the mass, L becomes a function of the cloud's mass and temperature ($L = f(M, T)$). In the following analysis $L_{\rm IR}/M_{\rm CO}$ will be considered and denoted as L/M, unless the use continuum-derived mass is specified explicitly.

To check for potential biases in the SFE-solenoidal-fraction relation, the signal-to-noise ratio and field size are considered. A negative correlation (Spearman $r = -0.27$, p-value $= 5.08 \cdot 10^{-18}$) was found between the solenoidal fraction and the SNR (defined for each cloud as the square root of the quadrature sum of the SNR values at the voxels within the extracted cloud, Figure 6.6). The field size and the solenoidal fraction show a small positive correlation ($r = 0.19$ p-value $= 2.02 \cdot 10^{-9}$, Figure 6.7). An evaluation of the effects of these correlations on the solenoidal-fraction-SFE relation through partial correlation analysis shows that none of these factors significantly impacts the negative correlation between the solenoidal fraction and the SFE (accounting for the SNR returns $r = -0.25$ with p-value $= 8.86 \cdot 10^{-15}$, while accounting for the SNR yields $r = -0.27$ with p-value $= 3.12 \cdot 10^{-17}$), nor does their combined effect ($r = -0.17$, p-value $= 1.3 \cdot 10^{-7}$).

A prominent feature of the plot in Figure 6.5 is the scatter that characterises the relation between SFE and the solenoidal fraction. The scatter appears small at low solenoidal fractions, increasing at the high solenoidal end. The 16 clouds with solenoidal fraction < 0.12 that populate the upper left corner of Figure 6.5) include both compact cores (150-600 voxels) and small clouds (1000-3000 voxels). Their average velocity dispersion is 1.5 $\mathrm{km\,s^{-1}}$. These clouds do not present special size-related qualities but can be considered as a sample of the full distribution as can be proven both by visual inspection and a Kolmogorov-Smirnov test (projected size: $k = 0.14$ with p-value $= 0.93$; linewidth: $k = 0.37$ with p-value $= 0.04$).

This change in the observed scatter may be a real feature of the L/M distribution or

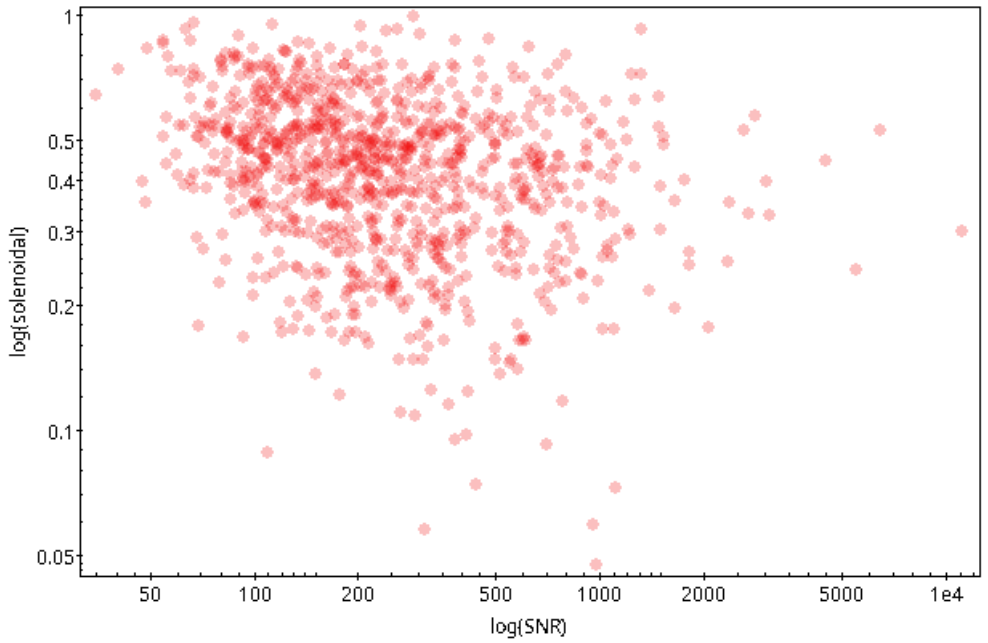


FIGURE 6.6: Solenoidal fraction as a function of the signal-to-noise ratio.

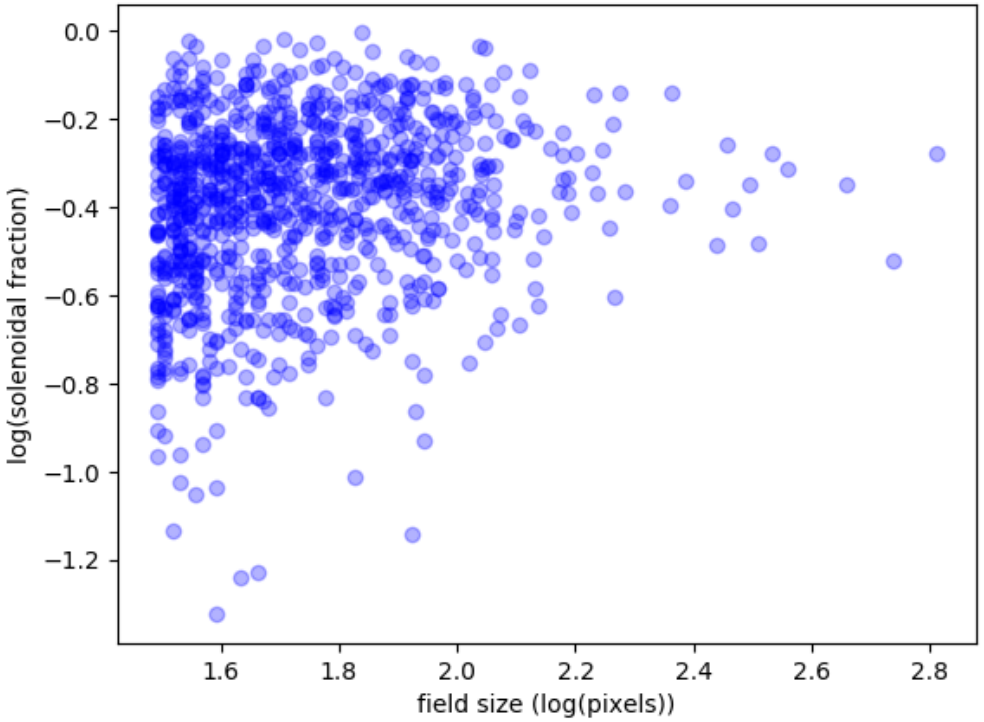


FIGURE 6.7: Solenoidal fraction as a function of the field size.

just be due to larger sample sizes revealing the wings of the distribution. The scatter at the high end is similar to the range seen in results by Eden et al. (2015) and Rigby et al. (2016) using the same L/M parameter. To understand if the scatter is related physical effects on L/M, such as the evolution of the IR emission, it should be compared to measurement uncertainties. Considering masses measured from the continuum, L/M becomes correlated with temperature and can be interpreted as an evolution indicator (Urquhart et al., 2018). The continuum clump masses may be related to cloud CO masses evidencing that there may be an evolution factor in the scatter in the SFE-solenoidal fraction relation. Evolutionary effects can then be factored out by considering the dust temperatures of the YSOs.

6.3 Scatter and temperature

Figure 6.8 shows the solenoidal fraction-SFE scatter plot centred around its weighted linear fit (red solid line in Figure 6.5). The weights of the fit correspond to the standard deviations of the distributions of values of SFE obtained after binning the solenoidal fraction. To investigate the scatter around this simple linear model, the Hi-GAL bolometric temperatures (colour-coded in Figure 6.8) are used. The bolometric temperature is defined from the flux density F_ν (Myers & Ladd, 1993) as

$$T_{\rm bol} = 1.25 \times 10^{-11} K \times \frac{\int_0^\infty \nu F_\nu d\nu}{\int_0^\infty F_\nu d\nu}. \tag{6.2}$$

The temperature associated with each SCIMES cloud corresponds to the average of the temperatures of the Hi-GAL sources it contains. In general, typical bolometric temperatures found in Hi-GAL clumps range from ~ 10 K (pre-stellar sources) to ~ 80 K. There is a positive correlation between the luminosity of the embedded massive protostars and the continuum temperatures of the gas clumps in which they were formed (Urquhart et al., 2011). Urquhart et al. (2018) extended this analysis to lower luminosity and less-evolved sources (pre-stellar), showing that, in the ATLASGAL sample, L/M is strongly correlated with the bolometric temperature of the source, which allows for the reliable prediction of one quantity, if the other is known. The authors also showed that the L/M-temperature relation holds over almost 6 orders of magnitude in L/M clump and the whole range of ATLASGAL temperatures.

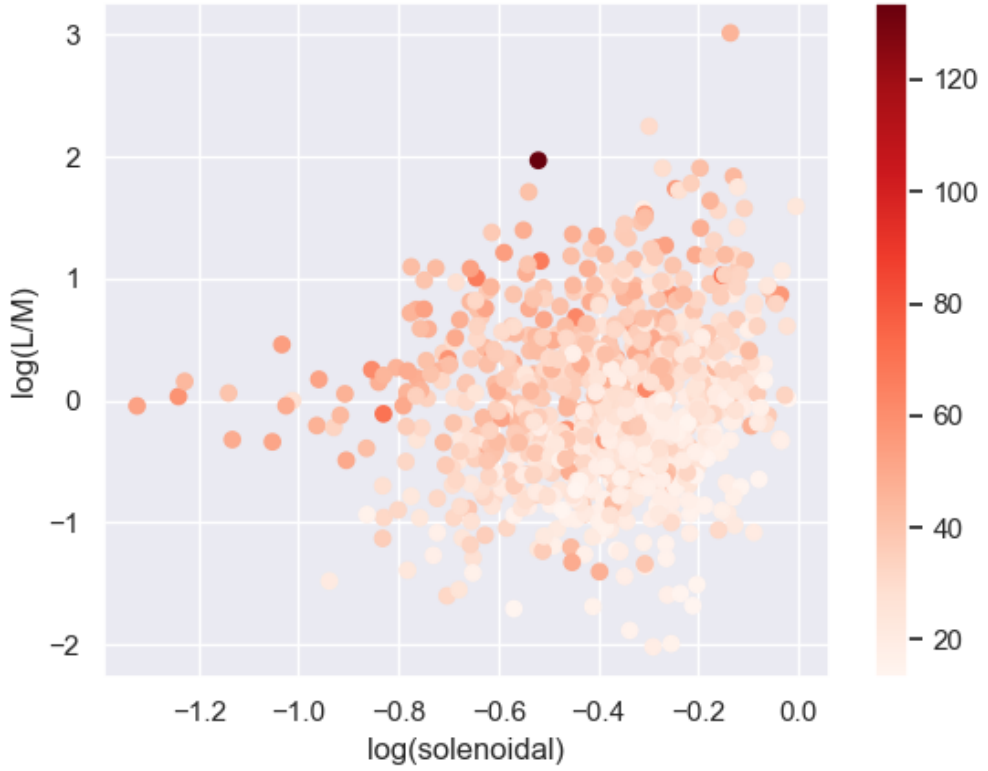


FIGURE 6.8: Adjusted scatter plot of the SFE and solenoidal fraction. The plot is centred around the weighted linear model shown in Figure 6.5. Colour coding corresponds to the Hi-GAL bolometric temperature associated with each source. Luminosities and masses are given in units of $L_{\odot}$ and $M_{\odot}$ respectively.

Furthermore, Urquhart et al. (2018) found that both luminosity and L/M are correlated with the dust temperature, but the large scatter in the data and the strong power-law relationship of the luminosity–temperature distribution make it difficult to use dust temperature as a measure of stellar evolution. On the other hand, the correlation between L/M with its lower power-law relation to temperature makes it a less sensitive parameter to small changes in temperature, Similar results were found independently by Elia et al. (2017) using the Hi-GAL sample.

There is no obvious correlation between the excitation temperature in the present data with independent CO masses, suggesting that the column density does not evolve significantly during the star formation process [1].

[1] Urquhart et al. (2018) tested this correlation for the ATALSGAL sample, finding that the column density decreases as the cloud evolves, however, they noticed that the weak correlation found may arise from an observational bias: the reduced sensitivity to lower column densities.

The adjusted scatter plot in Figure 6.8 displays a sharp increase of scatter in the SFE for solenoidal fractions $> 10^{-0.9}$. To check that the distribution of $log(L/M)$ at log(solenoidal) < -0.9 is statistically consistent with the distribution at log(solenoidal) < -0.9, a Kolmogorov-Smirnov test is performed over the two distributions. As above, the null hypothesis that the two samples (distributions) are drawn from the same distribution, while the alternative is that they are independent. With $k = 0.13$ and p-value $= 0.24$, the null hypothesis cannot be rejected and the log(SFE) distribution must be considered statistically consistent over log(solenoidal), i.e. the scatter is not a function of solenoidal fraction.

To quantify and filter out the scatter in L/M that may be due to temperature and, hence, evolution variations, the following steps are taken. First, we select a temperature bin (5-K wide) whose distribution of L/M approximates a normal distribution. This distribution is used as a filter to deconvolve the Gaussian that approximates the full L/M distribution. This method is illustrated in Figure 6.9.

The L/M ratio is independent of distance, so the uncertainty associated with it equals the quadrature sum of the uncertainty in the flux and the mass. Assuming the uncertainty depends only on the uncertainty of the column densities, it is about $\sim 20\%$ (Rigby et al., 2019).

The bolometric flux of a Hi-GAL source is evaluated as the sum of the areas of trapezia defined by flux values of consecutive bands (see Eden et al., 2012). The bolometric flux of a SCIMES cloud is then the sum of the fluxes of the HigGal sources it contains. For the errors of the HiGal bolometric fluxes, the fractional errors are obtained by summing the errors (quadrature sum of errors in the five wavebands) and dividing by the sum of the fluxes of the bands. This fractional error multiplied by the value of the bolometric flux of the source gives the error in the source's bolometric flux. The errors in the bolometric fluxes within a SCIMES cloud are again summed in quadrature to obtain the error associated with the cloud. This calculation yields an average error in the bolometric flux $\sim 7\%$ [2]. Estimating the percent variation coefficient of the deconvolved Gaussian distribution (variation coefficient, $c_v = 100 \times \sigma/\mu$) and converting it back to

[2]Notice that error in the bolometric flux is derived through the quadrature sum of the error at the five Hi-GAL wavelengths. Using a small number of wavelengths to estimate the error over the entire spectrum produces a lower value of the error. Thus one could say that the value from the Hi-GAL wavebands is a lower bound of the error in the bolometric flux.

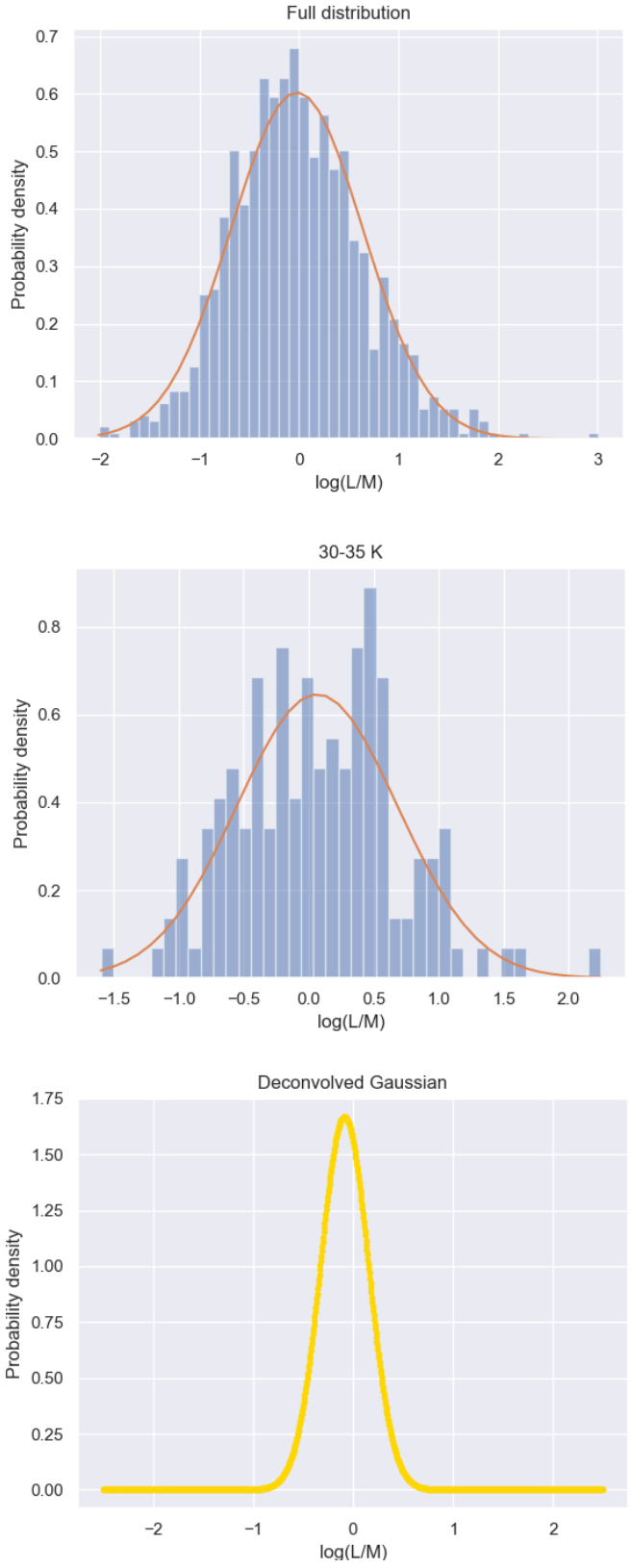


FIGURE 6.9: Deconvolution of the full L/M distribution by the Gaussian approximating the distribution in the $30 - 35$ K bin.

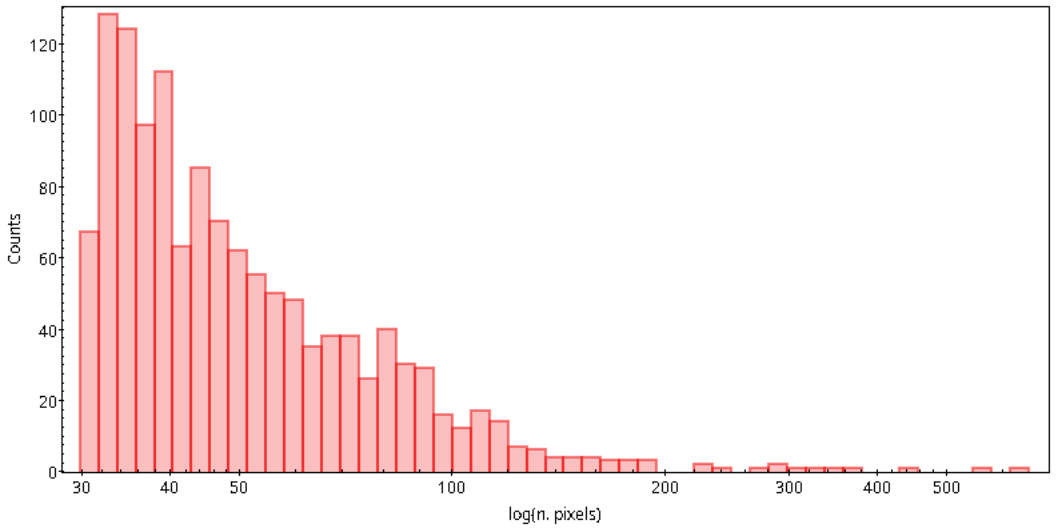


FIGURE 6.10: Distribution of field sizes including padding for the sample of CHIMPS clouds used in calculation of the solenoidal fraction.

the linear scale gives $c_v \approx 56\%$, this value suggests that the observed scatter does not originate from measurement errors alone, but other physical factors must be involved.

Different star formation efficiencies at similar values of the solenoidal fraction may thus be linked to different evolutionary stages of the clouds in the sample considered. The relation between bolometric luminosity and envelope mass indicator of the evolutionary status of a core/clump. $L_{\rm bol}$ versus M diagrams are widely used to trace evolutionary tracks of clouds (Saraceno et al., 1996; Molinari et al., 2008; Elia et al., 2013; Ragan et al., 2013). Evolutionary tracks are fundamentally divided into an accretion phase followed by a clean-up phase (Molinari et al., 2008; Smith, 2014). In the earliest stages of star formation, these tracks are nearly vertical as the YSO accretes mass from the surrounding envelope increasing its luminosity. When the central star reaches the zero-age main sequence (ZAMS), and the dispersal of the residual clump material begins, the track flattens into a nearly horizontal line. In a sample of clouds with different characteristics and located in different Galactic environments, clouds with similar fractions of solenoidal modes, may be at different stages of their evolution, manifested through the parameter after deconvolution with the temperature distribution L/M. This framework would explain the scatter observed in Figures 6.5 and 6.8.

6.4 The field size

The field sizes associated with the sample of clouds consider in the analysis presented above range from 30 to 649 pixels in side (see Figure 6.10). As considering the Fourier transforms on fields of small sizes (see Appendix C) is not likely to yield useful information on the state of the turbulence in the corresponding clouds, the validity of the results presented above was tested on a series of sub-samples with fields of decreasing size. This test showed that the results (distribution of solenoidal fraction with Galactocentric distance, the negative correlation between SFE and the solenoidal fraction, the scatter in solenoidal fraction-SFE plots) still hold when a sample of clouds with a field larger or equal to 85 pixels is considered. The sample size in this threshold case is reduced to less than 200 clouds. Above this threshold, the size of the sample is reduced drastically which invalidates the outcome of the analysis presented in Sections 6.1 and 6.2.

6.5 Summary

The computation of the solenoidal fraction was performed on a selected sample of molecular clouds (1311) in the SCIMES segmentation of CHIMPS. This analysis produced four main results:

- the solenoidal modes of turbulence appear to be higher in the inner Galaxy (although the sample in question only contains a small number of clouds associated with these distances),
- the solenoidal fraction declines with a shallow gradient with increasing Galactocentric distance,
- star formation efficiency and the solenoidal fraction are negatively correlated (which is consistent with the hypothesis that solenoidal modes prevent or slow down the collapse of dense cores),
- the significant scatter in SFE-solenoidal-fraction plots appears to be caused by physical factors such as different stages of cloud evolution.

Chapter 7

Discussion and conclusion

7.1 Fellwalker and SCIMES

The study of molecular emission in position-position velocity (PPV data) has been approached through a wide range of analytic methods. These techniques make use of different features of molecular emission to identify gas structures as discrete sets of connected voxels (segmentation) with emission (brightness temperature or column densities) above a specified threshold. Further selection criteria may then be employed to characterise these 'clusters' as individual molecular clouds allowing for the construction of a consistently-selected set of "objects" which can be used for statistical studies of cloud properties, star-formation, and chemistry. The entire segmentation process is performed with a variety of automatic algorithms (see Section 1.4). However, as the ISM is a continuous medium, the discrete segmentation of the emission is bound to introduce artificial structures independently of how physically realistic and sophisticated the chosen paradigm is. Such segmentation may thus be more suitable for the power-spectrum-like analyses of the continuous data (Eden et al., 2021). These extraction methods are often complex and it is difficult to compare their relative efficacy or quantify their biases since their core algorithms are based on widely different paradigms and few have been applied to the same data. Furthermore, there is no commonly used standard against which these techniques are calibrated. From this standpoint, it is interesting to set up a direct comparison between different segmentation methods by applying them to the same data and with a suitable choice of input parameters. Finding matching values of the input parameters for different segmentation algorithms may not always be

possible, especially if the methods are based upon utterly different principles. Such a choice also require knowledge of an optimal parameterisation for each method on the given data set.

Chapter 4 presented an attempt to cross-correlate the properties of individual clouds in two different segmentations of the ^{13}CO (3-2) emission in the CHIMPS survey: one obtained with the watershed algorithm FellWalker and the other with the dendrogram based SCIMES (Chapter 3). SCIMES is a recent dendrogram-based image segmentation method that uses clustering theory to identify emission sources. In this framework, clustering is not established by the proximity of neighbouring pixels, but through similarity criteria based on the physical properties of the molecular gas (see B). This defining characteristic of SCIMES mitigates the influence of survey-sensitivity biases. The FW algorithm, on the other hand, is a variation of the watershed paradigm. It extracts emission structures locally, through the path of steepest ascent around local emission peaks. These methodologies yield different numbers of molecular clouds (SCIMES 2944, FW 3665) but produce largely consistent results with similar ranges in masses ($M/M_{\odot} \simeq 10^{0.4-5.0}$ and $M/M_{\odot} \simeq 10^{0.6-4.8}$), sizes (no.voxels $\simeq 10^{1.5-5}$ and no.voxels $\simeq 10^{1.8-4.2}$), equivalent radii $R_{\rm eq}{\rm pc}^{-1} \simeq 10^{-0.7-1.3}$ and $R_{\rm eq}{\rm pc}^{-1} \simeq 10^{-0.6-1.0}$), mean number densities ($\overline{n}H_2/{\rm cm}^3 \simeq 10^{0.9-3.7}$ and $\overline{n}H_2/{\rm cm}^3 \simeq 10^{1.6-4.0}$), and velocity dispersions ($\sigma_v/{\rm kms}^{-1} \simeq 10^{-0.55-1.0}$ and $\sigma_v/{\rm kms}^{-1} \simeq 10^{-0.53-0.7}$). The distributions of the quantities investigated: mean number densities (Figure 4.15), masses (Figure 4.11), the virial parameters (Figure 4.31), and dynamic timescales (Figures 4.20 and 4.19) all reflect the differences in volumes and geometries found in the two segmentations (Figures 4.9 and 4.8).

Additionally, the SCIMES extraction for the ^{12}CO (3 - 2) in COHRS is considered as a term of comparison with a different tracer over the same area spanned by CHIMPS. This particular transition of ^{12}CO isotopologue is, in general, a more optically thick tracer than ^{13}CO (3 - 2). In practice, this implies that COHRS segmentation traces lower density regions of the molecular clouds, that are not detected in CHIMPS. The linewidths for the COHRS clouds will thus be naturally wider than those found through both SCIMES and FW (section 4.4.4). Probing lower-density emission, COHRS detects larger structures than CHIMPS. The inconsistent results in the SCIMES segmentations of ^{12}CO and ^{13}CO emission can be traced back to the ^{12}CO abundance and optical of

the depth of the isotopologues as well as to the different SCIMES parameterisations chosen for the segmentations.

A closer look at the distribution of the assigned SCIMES heliocentric distances (Figure 4.3) and the independently generated Galactocentric distances (Figures 4.6 and 4.7) reveals that both distributions display the same features as the FW assignments. The difference in distance assignment has supposedly little influence on the distance-dependent physical properties. Size–linewidth (Figure 4.22), size–density (Figure 4.16), and size–virial parameter (Figure 4.32) plots for the CHIMPS clouds, also reveal similar relations. An identical situation is reported by (Lada & Dame, 2020) in their studies of mass-size relations (Larson, 1981) and the GMC surface densities in Galactic clouds. Lada & Dame (2020) compared data from the SCIMES (Rice et al., 2020) and FW (Miville-Deschênes et al., 2017) extractions of ^{12}CO in the low-resolution CfA-Chile survey (Dame et al., 2001). The mass-size relation they found did not appear to be particularly sensitive to differences in the two methodologies used for the emission segmentation.

The distributions of velocity dispersions (Figure 4.21 and excitation temperature (Figure 4.23) only depend on the size of the clouds as identified by each algorithm (number of voxels that constitute a cloud). The SCIMES extraction includes both smaller and larger sources than FW (see Figure 4.9). The size comparison presented in table 4.1 suggests size and number of clouds extracted by the two algorithms depend on the environment. In crowded areas (large star formation complexes like W43 ($l = 30.8°$, $b = 0.0°$)) a FW tends to split clouds into smaller clumps. Visual inspection reveals that the FW clumps have touching sharp borders (see Figure 4.1) whereas SCIMES identifies a single structure. The introduction of artificial boundaries between emission peaks is a consequence of the watershed algorithm which characterises disjoint clouds by single individual peaks. This method "cuts the valleys" between peaks into separate assignments, thus splitting the envelopes of more rarefied structures enclosing denser clumps. This defining characteristic makes FW and similar methods better suited to extract sources in less crowded fields or to identify compact cores in crowded fields through a careful selection of the configuration parameters.

With the chosen parameterisation, SCIMES, on the other hand, register such structures as part of a single entity, thus proving to be more sensitive to tenuous emission in complex

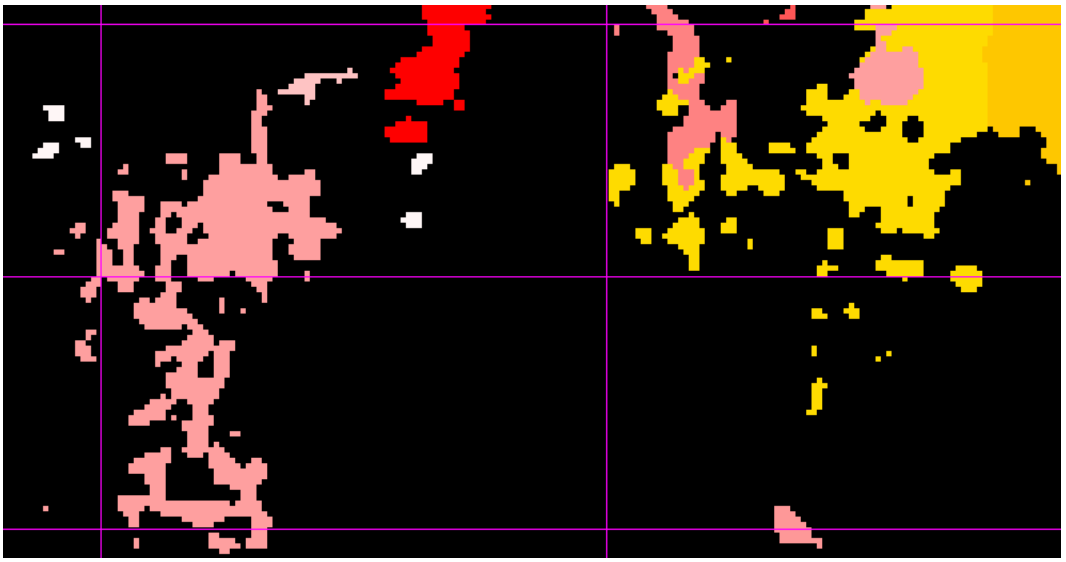


FIGURE 7.1: Example of disconnected clouds in the FW segmentation. The panel show the projection along the spectral axis of a portion of the FW extraction. The colors indicate individual clouds.

gas distributions and crowded fields. A number of cases of disconnected emission that is labelled as the same cloud emerged from inspecting the projection along the spectral axis of FW clouds (see Figure 7.1). The projection of some FW clouds also misses parts in their interiors (holes). These features are not present in the SCIMES extraction. Establishing a relationship between the results of the two methods requires the accurate analysis of substructures in individual clouds in different environments. This would allow for the identification of FW clouds within the SCIMES dendrograms, matching them with branches and subbranches[1] in the dendrogram hierarchy.

The differences in morphology and density observed in the SCIMES catalogue originate from SCIMES being, by definition, more sensitive to the global distribution of gas encoded as a single dendrogram. Consequently, the parametrisation that defines SCIMES dendrograms, which was chosen to match the FW configuration used by Rigby et al. (2019), has a significant impact on this study (see Colombo et al., 2019; Duarte-Cabral et al., 2021).

Furthermore, if an algorithm produces over-segmentation of molecular gas the total data sum in each clump (for instance the sum of the emission values associated with the voxels in the cloud) will, on average, be too low. The performance of an algorithm with respect to over-or under segmentation can be evaluated through the distribution of the measured

[1] In general, SCIMES leaves were found to be smaller than FW smallest clumps!

total data sum in each cloud compared to the expected distribution constructed from a set of artificially generated clouds. If a set of identical clouds (same peak amplitude and size) is considered, the distribution of measured total data sums will peak at the expected value, but will always have a tail of higher-valued clumps due to the random spatial positioning of clumps causing some clumps to overlay each other. An optimal clump-finding algorithm should not produce any significant tail of lower-valued clumps.

7.2 Analysis of turbulence

A potential driving agent of star formation has been identified as the relative fraction of turbulence modes in the interstellar molecular gas. Any connection between the properties of molecular clouds and their environment would show a dependence on galactic dynamics and/or the history of individual cloud formation (see section 8.1). Specifically, this involves both volumes and maximum mass scales (Hughes et al., 2013; Reina-Campos & Kruijssen, 2017) and physical properties such as cloud surface and volume densities (Sun et al., 2018), turbulent pressure and velocity dispersion (Heyer et al., 2009; Field et al., 2011; Shetty et al., 2012; Kruijssen & Longmore, 2013), and virial parameter (Sun et al., 2018; Schruba et al., 2019). Theoretical models and observations have demonstrated that these properties are correlated to star formation rate, and cluster formation efficiency, which typically increases with the gas pressure in the galactic plane (Vàzquez-Semadeni, 1994; Krumholz & McKee, 2005; Elmegreen, 2008; Padoan & Nordlund, 2011; Kruijssen, 2012; Adamo et al., 2015).

Molecular clouds form through the condensation of the lower-density ISM gas, thus inheriting its turbulent and shear-driven motions (Meidt et al., 2018, 2019; Kruijssen et al., 2019b). Galactic dynamics can thus stabilise clouds (Meidt et al., 2013) or compress them promoting star formation (Jeffreson & Kruijssen, 2018). This mechanism leads to the formation of shock-bounded layers between convergent flows, a process that induces fragmentation through non-linear instabilities (Vishniac, 1994). Numerical simulations of this scenario (Hunter et al., 1986; Klein & Woods, 1998) show that fully developed turbulence arises in the shock-driven layers (Hunter et al., 1986; Klein & Woods, 1998). This turbulent state is maintained throughout the duration of stream collision and its fragmentation into molecular clouds. The internal turbulence of molecular clouds originates from a dissipative energy cascade in compressible turbulent flows. At every scale,

the fraction of the energy that is not dissipated through shocks is transmitted to smaller-scale structures (Kornreich & Scalo, 2000).

In this framework, the high star formation efficiency (SFE) observed in disc clouds is linked to the prevalence of compressive (curl-free) turbulent modes. In contrast, the low SFE that characterises clouds in the Central Molecular Zone (CMZ) is related to the shear-driven solenoidal (divergence-free) component. A similar analysis of the Orion B molecular cloud (Orkisz et al., 2017) finds that the turbulence is mostly solenoidal, consistent with the low star formation rate associated with the cloud. The forcing is however position-dependent and varies with scale within the cloud with motions around the main star-forming regions being strongly compressive. Thus, this significant inter-cloud variability of the compressive/solenoidal mode fractions may be a decisive agent of variations in the SFE. Chapter 6 collects the results of the first full-sample study of turbulent modes in CHIMPS molecular clouds with a focus on their relation to star formation efficiency.

A software package capable of automating the calculation of the solenoidal fraction for a large sample of molecular clouds was designed and developed from the recipe described in section 5.2. This package produces the value of the solenoidal fraction, given a cloud map, emission data, and column density data as input and choosing a fitting model for the one-dimensional power spectrum. Further development of this tool is underway, and it is going to be used for several projects related to the investigation of turbulent modes in interstellar gas. The computation of the solenoidal fraction was performed on a selected sample of molecular clouds (1311) in the SCIMES segmentation of CHIMPS. This analysis produced two main results:

- the relative power in the solenoidal modes of turbulence appears to be higher in the inner Galaxy (distances < 4 kpc from the centre). The solenoidal fraction then declines with a shallow gradient with increasing Galactocentric distance. If confirmed by the analysis of a sample at lower longitudes, this result would be consistent with the disc becoming stable against gravitational collapse and the star formation rate being suppressed by the influence of the rotation of the Galactic bar;

- there is a negative correlation between star formation efficiency and solenoidal fraction consistent with the hypothesis that solenoidal modes prevent or slow down the collapse of dense cores (Figure 6.5).

These findings agree with the variation of SFE with the Galactic environment measured using both the numbers of HII regions per unit molecular gas mass and the dense gas mass fraction (DGMF). The DGMF peaks at around 3–4 kpc and then decline in the inner zone (Eden et al., 2012, 2013), where the disc becomes stable for the life of the bar (James & Percival, 2016). Two mechanisms are currently believed to cause the quenching of star formation in the regions around the bar. One theory focuses on the shock and shear arising from the rotation of the bar. The turbulence they induce in the molecular gas in the region stabilises clouds against collapse and thus inhibits star-formation. This scenario holds under the assumption that during its formation the bar collects most of the gas in the central regions within the co-rotation radius (Tubbs, 1982; Reynaud & Downes, 1998; Haywood et al., 2009; Khoperskov et al., 2018). While the second mechanism is identified with the torque generated by the rotation of the bar. This force induces gas to migrate from the Galactic outskirts to the central regions. This inflow fuels nuclear star formation but deprives the regions close to the bar of gas, thus suppressing star formation (Spinoso et al., 2017). Kiloparsec scale formation "deserts" were observed at the centre of barred galaxies (James et al., 2009) [2].

Outside the Inner Galaxy, the solenoidal fraction shows a negative correlation to distance (for Galactocentric distances greater than 4 kpc, a Spearman test returns $r = -0.133$ with p-value $= 1.498^{-6}$) and declines with a shallow gradient with a slope of -0.02 with no signal present at the spiral-arm radii. This result is in agreement with previous studies that found no significant arm associated signal in the fraction of star-forming compact sources (Ragan et al., 2016, 2018, and section 1.6). These findings suggest that the solenoidal fraction is unaffected by large scale features such as radial variations in density, shear, and metallicity and that differences between the individual clouds are more relevant to star formation. This picture challenges the idea that spiral arms may be direct triggers of star formation and considers them as mere sources of source crowding (Moore, 2012; Ragan et al., 2016). The increased star formation observed in

[2]The stellar populations observed ranged between $250 \cdot 10^6$ to $250 \cdot 10^9$ years (James & Percival, 2015a,b). Star formation deserts have not been found for older populations. These results strengthen the link between the star formation properties of central regions and the life cycles of the Galactic bar.

the spiral arms may be a consequence of their function as organising features that affect the ISM by delaying and crowding the gas that traverses them Dobbs et al. (2011). Spiral arms thus enable longer-lived and more massive molecular clouds. The longer lifetimes of molecular clouds in turn result in longer star formation time scales and consequently an increased SFE compared to inter-arm gas (Roman-Duval et al., 2010). In larger and denser clouds, the column density of clouds affects the mass function of massive clusters (Krumholz et al., 2009). Radiative heating in high-column density clouds suppresses fragmentation but does not appear to influence the clouds' overall SFE (Krumholz et al., 2010). Spiral arms are also likely to differ from one another (Benjamin et al., 2005). The inner and outer segments of the same arm may also impact star formation in different ways. The entry shock that the ISM gas undergoes upon entering a spiral arm is supposed to only exist within the corotation radius. This is the distance at which there is a differential velocity between the spiral pattern speed and the orbital rotation speed of the galactic ISM [3]. Outside the corotation radius, the SFE (and in general the state of the ISM) is expected to be governed by supernovae (Kobayashi et al., 2009; Dib et al., 2009). Krumholz et al. (2009) also predicts that internal radiative feedback dominates molecular gas (in non-starburst conditions).

Star formation declines abruptly in the Central Molecular Zone (CMZ) within 0.5 kpc of the centre (Longmore et al., 2013; Urquhart et al., 2013). The CMZ presents the highest abundance ($\sim 100\%$) of molecular gas in the Galaxy. The amount of molecular gas declines with increasing Galactocentric distance to only a few per cent at radii greater than 10 kpc (Sofue & Nakanishi, 2016). The inner Galaxy, particularly the CMZ, is a key environment to investigate SFE, but are not covered by CHIMPS and information from different surveys is therefore required to probe these environments (see Chapter 8).

In this thesis, a negative correlation between the solenoidal fraction and SFE defined as $L_{\rm bol}/M$ (see section 6.2) was confirmed. A prominent feature of the SFE-solenoidal fraction relation shown in Figure 6.5 is the large scatter that characterises the plot. Section 6.2 shows that this feature remains after deconvolution with a Gaussian approximating the variation in bolometric temperature representing the evolution of the individual embedded sources. The spread of the deconvolved distribution is still larger than the estimated errors in the SFE (derived from the errors in the fluxes in the Hi-GAL catalogue). This remaining scatter seems to arise from physical factors linked to the state

[3] (to be just beyond the solar circle in the Milky Way Lépine et al., 2011)

of the cloud and its evolution. This conclusion is emphasised by the relation of the L/M, the parameter used to define SFE, to the evolution of clumps (see discussion in section 8.3). The analysis presented in Urquhart et al. (2018) reveals trends for increasing temperatures and luminosities with the evolutionary stage of the embedded stars as they advance towards the ZAMS stage. Changes in both L and M can be attributed to feedback from the forming protoclusters on their natal clump. These variations due to stellar feedback are reflected in the linewidths of molecular transitions (Urquhart et al., 2018).

The value of the solenoidal fraction assigned to each cloud accounts for the overall modes of the gas it contains, with substructure contributing over all spatial frequencies. Thus, the same value of the solenoidal fraction can be attained through different configurations of molecular gas, i. e. different cloud sizes, velocity distributions, densities, amount of molecular gas, number, and size of star-forming cores, and stellar feedback mechanisms. Although compressive turbulence remains one of the driving agents of star formation in this framework, star-forming regions can be affected by several factors that slow down their collapse. In addition to the delay induced by the thermal pressure gradient at early stages of collapse, magnetic fields (even if the clouds are magnetically supercritical, i.e. the magnetic energy is less than the binding energy, Inoue & Inutsuka, 2012; Vàzquez-Semadeni et al., 2011; Girichidis et al., 2018), Galactic differential rotation through shear and Coriolis forces (Dobbs & Baba, 2014; Meidt et al., 2020), and the non-spherical (planar or filamentary) shape of the clouds (Toalà et al., 2012; Pon et al., 2012) contribute to delaying collapse. If the magnetic support is weak, star formation is expected to proceed more efficiently and star clusters can be formed. For clustered star formation, numerical simulations show that stellar feedback such as protostellar jets, outflows, and stellar winds can inject supersonic turbulence in molecular clumps (Nakamura & Li, 2007; Offner & Arce, 2015), and the clumps can be kept near virial equilibrium for several dynamical timescales.

Chapter 8

Future work

This thesis initiated a full-sample study of turbulent modes in Galactic molecular clouds. The investigation explored the relationship between the solenoidal fraction and star formation efficiency in the CHIMPS survey and hinted at a gradient in solenoidal modes extending out from the inner Galaxy. Along with a fully developed software package for the automated calculation of the solenoidal fraction over large samples of clouds, this thesis naturally sets the foundations for the extension of the statistical analysis of turbulent modes and SFE to different Galactic environments and a selection of individual clouds. This analysis on high-resolution surveys could also shed light on the factors behind the scatter appearing in the solenoidal-fraction-SFE relation.

8.1 Turbulence in different Galactic environments

A primary objective is to extend and improve the statistical analysis of turbulence initiated in this thesis with the aim to link solenoidal modes at different Galactocentric distances and over a wide range of scales to both SFE and other physical (temperature, density) and geometric properties (shape factor, internal structure of the dendrogram) of clouds, clumps, and cores. Particularly interesting is the estimation of the solenoidal fraction in filamentary structures since these features appear to host the majority of star-forming cores (Polychroni et al., 2013; Könyver et al., 2015). This study would however require to ascertain at what scales the loss of symmetry/isotropy within such structures affect the applicability of the method.

The investigation is going to focus on molecular gas in three key Galactic environments:

- the inner Galaxy within 3 kpc, where the disk is becoming stable against gravitational collapse, and the star formation is quenched by the rotation of the Galactic bar (see discussion in the previous chapter, section 7.2);
- the Central Molecular Zone (within 0.5 kpc) where star formation plummets while the molecular-gas fraction increases towards 100 %. The high turbulent energy in this region manifests as line-widths of $\sim$ 10–20 $\mathrm{km\,s^{-1}}$ on parsec scales (Henshaw et al., 2016). Such high turbulence raises the critical volume density threshold for star formation (Krumholz & McKee, 2005; Federrath & Klessen, 2012) may explain the difference between the SFR predicted by density thresholds and the current SFR in the region (Lada, 2010; Lada et al., 2012). Recent high-resolution surveys of the CMZ have evidenced the lack of internal structure in CMZ dense clouds (Battersby et al., 2020; Hatchfield et al., 2020) indicating that the formation of such structures is impeded by the highly turbulent environment.
- the outer Galaxy beyond a radius of 10 kpc, where the molecular fraction drops to a few per cent, molecular clouds are sparsely distributed (Wouterloot et al., 1990) and the metallicity (Rudolph et al., 1997), the diffuse Galactic interstellar radiation (Bloemen, 1985), and cosmic-ray flux density (Bloemen et al., 1984) is reduced compared to the Solar neighbourhood. Outer Galaxy clouds were also observed to be less massive than clouds found in the Inner Galaxy (Brand & Wouterloot, 1995). In general, they possess larger radii than their equally massive Inner Galaxy counterparts (Brand et al., 2001). The reduced pressure of the surrounding ISM at large Galactocentric distances is thought to account for these observations. This environment allows for the study of the influence of the reduced pressure on cloud formation/turbulence, and SFE.

To cover these regions a combination of data from different surveys and tracers is required. The extensive Structure, Excitation and Dynamics of the Inner Galactic Interstellar Medium survey (SEDIGISM, Duarte-Cabral et al., 2021), covers 78 $\mathrm{deg^2}$ of the inner Galaxy ($60° \leq l \leq 18°, |b| \leq 0.5°$) in the $J = 2 \rightarrow 1$ rotational transition of ^{13}CO. This survey provides a detailed, global view of the inner Galactic interstellar medium at a resolution of $\sim$ 30". In addition, the following surveys are considered:

COHRS (Colombo et al., 2019, see section 2.2), and CHIMPS 2 (Eden et al., 2020), the ongoing extension of CHIMPS. This ^{12}CO / ^{13}CO / C^{18}O ($J = 3 \rightarrow 2$) survey extends CHIMPS and COHRS spanning Galactic longitudes between 28° and −5°, thus probing the innermost 3 kpc of the Galaxy including the CMZ in its entirety. Integration with high-resolution ALMA observations of the CMZ clouds (especially in the light of the recent confirmation of star formation in G0.253+0.016) will also be analysed with methods developed here (Walker, 2021). The high resolution of the ALMA dataset should enable an analysis of the internal turbulent structure (assuming the necessary apodisation of the maps).

The outer Galaxy portion of CHIMPS 2 spans longitudes between 215° and 225°. In this region which is also included in the FUGIN ($J = 1 - 0$, Umemoto et al., 2017) and Hi-GAL surveys (section 2.4) and contains over 1000 star-forming and pre-stellar clumps (Elia et al., 2013). The Forgotten Quadrant Survey (Benedettini et al., 2020) also covers this sector in ^{12}CO and ^{12}CO ($J = 1 \rightarrow 0$).

The construction of this extensive catalogue linking solenoidal modes to different Galactic environments and structural properties of the clouds will help shed light on both the impact of Galactic molecular environments on SFE and the not-so-well understood transitions between environments characterised by different abundances and densities of molecular gas (see section 6.2). In particular, the investigation will focus on the transition at the boundary of the CMZ which marks the onset of higher turbulent pressure and, consequently, a heightened density threshold for star formation (Kruijssen & Longmore, 2014; Sormani et al., 2019).

8.2 Selected clouds

Investigating the turbulent modes within restricted regions or at different scales within individual clouds in the various Galactic environments is also advantageous for quantifying the impact of the environment on the clouds' internal structure (and consequently their SFE). Of particular importance is the identification of cloud collisions or colliding neutral flows associated with enhanced compressive turbulence.

This approach has been applied to observations of the star-forming complexes in Orion B by Orkisz et al. (2017), who showed how the values of the solenoidal fraction increase

with scale, zooming out from the densest cores. Their method however involves the introduction of artificial boundaries to trace out the edges of the fields considered, with the potential addition of steep gradients in the emission (and/or velocity) distribution in these areas. In this situation, the treatment of boundaries becomes crucial to ensure that the assumptions that guarantee the applicability of the method developed by Brunt & Federrath (2014) are satisfied.

The study of turbulent modes within an isolated cloud is also a crucial tool to understand the role dense gas plays in regulating star formation efficiency (SFE). Dense gas ($10^4 - 10^6$ cm^{-3}, Lada, 2010) is vital to the star formation process (see 1.1 and 1.2) and higher critical density molecular-line tracers, such as HCN are excellent at probing the behaviour of the dense gas most closely associated with star formation (Onus et al., 2018). A recently accepted JCMT proposal for observation of clouds in HCN and HCO$^+$ $J = 4 \rightarrow 3$ within the Milky Way is aimed at investigating how the dense-gas SFE ($L_{\rm IR}/L_{\rm HCN}$[1]) varies across the Milky Way. In particular, one of our objectives is to test competing ideas that star formation is controlled by the free-fall time (Krumholz et al., 2012) or a dense-gas threshold (Lada et al., 2012). HCN data are also going to be used to investigate two individual star-forming regions. The most massive star-forming region in the Milky Way, W43, which is expected to have a high star formation rate in the future (due to its massive and dense areas (Motte et al., 2003)), but its current SFE is consistent with the rest of the Plane clouds (Eden et al., 2012). In contrast, the W49 region is statistically influencing global star-forming properties (Moore, 2012) and contains over 5% of the infrared YSO luminosity of the Galaxy (Urquhart et al., 2014a). By comparing the turbulent modes of dense gas to the ration$L_{IR}/L_{\rm HCN}$ in these two regions, it is possible to investigate the influence that dense gas has on star formation. The region W43 ($l = 30.8°$, $b = 0.0°$) is a precursor of a true mini-starburst system, while W49 ($l = 43.2°$, $b = 0.0°$) is one. The HCN kinematics will tell us more about the role of dense gas as a function of time.

The study of isolated clouds naturally extends to the investigation of the evolution of turbulent modes in artificial samples. In particular, snapshots of magneto-hydrodynamical simulations of the collapse of turbulent molecular clouds (Teyssier, 2002; Smith et al., 2020; Izquierdo et al., 2021) may shed light on the evolution of the solenoidal fraction in

[1]A denser gas tracer will give a different $L_{IR} - L_{gas}$ relationship from more diffuse gas as the free-fall time is shorter at higher densities. The ratio of $L_{\rm IR}/L_{\rm HCN}$ will test the dense-gas threshold theory as it should remain constant as the amount of star formation should scale with the amount of dense gas.

a controlled environment. Of particular interest is testing the limits of the applicability of the method in the presence of magnetic fields of increasing strength.

8.3 Scatter and clouds evolution

Finally, equipped with an extensive catalogue of sources spanning the critical Galactic environments and the information about the distribution of dense gas, a deeper analysis of the scatter in the SFE that occurs in Figures 6.5 and 6.3 can be performed. As it was shown in Chapter 6, this feature appears not to be caused by measurement errors but to arise from physical factors. The scatter is still prominent after deconvolution with the bolometric temperature, a proxy evolution parameter for the sources embedded in the clouds. The SFE measure adopted ($L_{\rm bol}/M$) is itself an indicator of the evolution of clouds (measured by their luminosity), which reinforces the assumption that clouds with similar solenoidal fractions may be at very different stages of their evolution. The value of the solenoidal fraction assigned to each cloud accounts for the overall modes of the gas it contains, with substructure contributing over all spatial frequencies. Profoundly different configurations of molecular gas (i.e. different cloud volumes, velocity distributions, densities, etc.) and may thus result in very similar values of the solenoidal fraction. Further analysis will focus on quantifying the amplitude of scattering and disentangle (with the aid of high-resolution data, i.e. ALMA, CfA) the factors that may produce it. This step will consider the amount of dense gas and the properties of the Galactic environment that hosts the clouds.

Appendix A

The FellWalker algorithm

The FellWalker (FW) algorithm implements a variation of the watershed paradigm. The segmantation perfomed by watershed algorithms consists in the identification of regions of catchment basins (areas of low emission) around local minima in the emission. The watershed lines that separate the basins constitute the boundaries of the low emission regions (Roerdink & Meijster, 2001). On the other hand, FW first searches for local maxima and partitions the dataset through gradient tracing by separating regions that correspond to the maximum values of the emission. The FW design aims to overcome the issues arising in algorithms based on the analysis contour levels. This class of algorithms considers a set of equispaced contours defined by two main parameters: a set baseline emission and the interval between adjacent levels. For three-dimensional data and crowded fields, the resulting segmentation becomes very sensitive to the spacing between contours. Choosing too large an interval might exclude real emission peaks, while an interval that is too small may cause noise spikes to be selected as true emission (Brunt et al., 2003; Elia et al., 2007; Smith et al., 2008; Kainulainen et al., 2009; Pineda et al., 2009). Finding a good compromise on the contour interval thus becomes crucial to the final decomposition. Moreover, in this framework, the segmentation is solely determined by those voxels that belong to the contour lines, a small fraction of the emission values contained in the datacube.

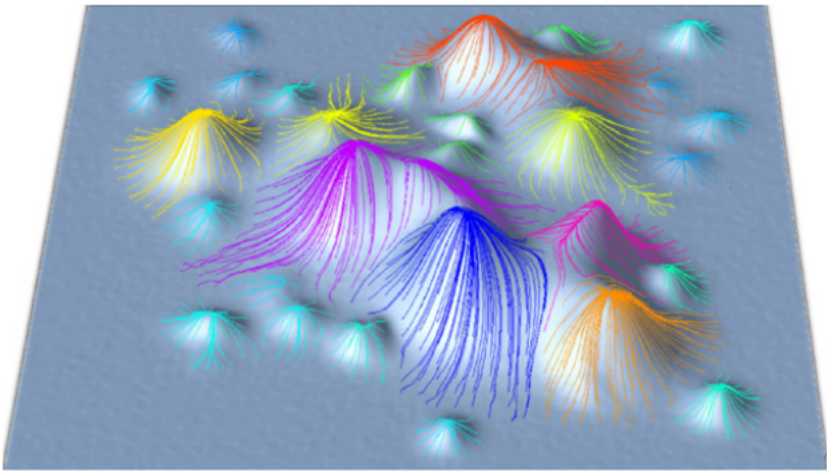


FIGURE A.1: Representation of the emission peaks found through paths of steepest ascent in the FellWalker algorithms. The emission cloud uniquely associated to each peak is highlighted in color. The 'landscape' emerging from this picture is reminiscent of the 'fells' of Northern England, hence the name 'FellWalker'. Figure reproduced after Berry (2015).

A.1 Algorithm

The FellWalker strategy determines the paths of the steepest ascent originating at each data point with an emission value that exceeds a given baseline threshold. It then uses the set of paths associated with the same peak to identify the cloud in the emission data array. A path of steepest ascent is a sequence of data points in which each successor is the nearest neighbour of the predecessor with a higher emission value than any point in the sequence so far.

A path is constructed by stepping from a voxel to its highest-emission nearest neighbour. The search is repeated at this new point. The sequence continues until a summit is reached: no point with higher emission values are found. At this point, FW looks for a voxel with higher emission in a larger neighbourhood. The size of this neighbourhood is determined by an input parameter. If a point with a higher emission value is found, then the path continues from this point. Otherwise, the path terminates. The union of all paths terminating at the same summit constitutes a cloud in the emission (see Figure A.1).

If a path meets a point that already belongs to a cloud, the path is terminated and its points are added to the cloud (see Figure A.2). Thus, given an emission array, FW

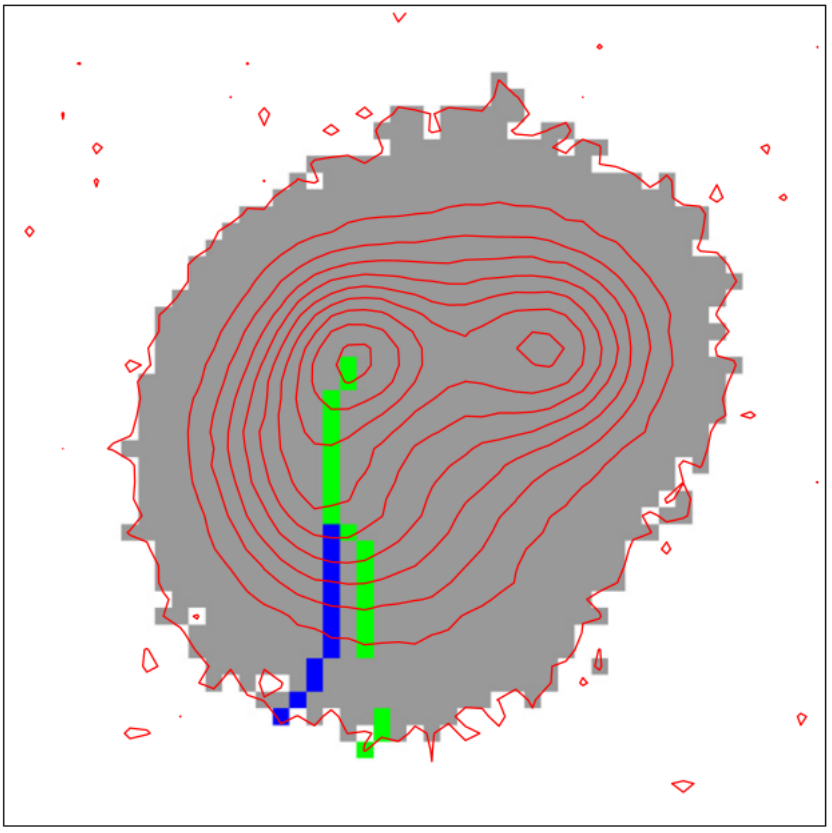


FIGURE A.2: Two paths of steepest ascent within an artificially generated emission cloud. The contours show the data values. Pixels above the baseline threshold are coloured grey, while pixels have values below the threshold. The green pixels form a path that terminates at the left peak. The blue pixels trace a path that was initiated after the green path and terminates at the pixel where the two paths cross. Thus both paths belong to the same cloud. Notice that after the first three pixels, the green sequence stops at a noise spike to continues at the highest emission value within a 9-by-9 neighbourhood. Figure taken from Berry (2015).

segments it into a number of disjoint subsets (clouds) characterised by single individual emission peak.

This feature of the FW algorithm makes it particularly well-suited for the identification of dense gas feature such as clumps associated to emission peaks, and thus the star-forming structures. The most reliable segmentation results are thus obtained when the method is used on the emission from isotopologues and transitions that trace denser regions of the molecular ISM (Roueff et al., 2020).

In practice, the operations described above are recorded in a clump assignment array (CAA), an array of integer values of the same shape and size as the emission data. Through masking and matching, voxels with emission values below the baseline are

labelled -1 in the CAA. Usable voxels are initially flagged with 0. Isolated 0's are re-labelled to -1. This step enables the identification and removal of isolated noise spikes from the final assignment. The core algorithm is then run at each emission voxels with the label 0 in the CAA. A unique identifier is issued for all voxels corresponding to an individual cloud. The CAA thus stores the positions and the identifiers of all clouds. The gradient of the ascent may vary greatly along different paths. Some paths start with very steep gradients, while in others a substantial ascent only occurs after a long section of low gradients. A path can be set to begin after a fixed minimum gradient is reached. The point in the sequence before this mark are discarded and not recorded in CAA. The new path is set to begin where the average of the gradient over four consecutive points of the original path exceeds this value (Berry, 2015). Applying this simple algorithm as it is to plateau regions may result in the extraction of well-distanced small clouds that differ only by small dips in the emission. This over-segmentation is resolved through the introduction of a parameter that specifies the minimum dip above which clouds are considered as separate entities. Clouds separated by 'emission valleys' below this value are merged into one single cloud. After merging the raw clouds, smoothing can be applied to mitigate the effects of the noise at the boundary between adjacent clouds (see below). This is achieved using a specified number of steps of a cellular automaton (one by default) to modify the integer values in the CAA. At each step, the cellular automaton produces a new CAA from the old one. Each entry of the new CAA is set to the most commonly occurring value in a 3-voxel sided neighbourhood of the point.

The final selection of clouds can be refined by excluding clouds that end at the edges of the emission array or clouds adjacent to areas of missing voxels. In addition, input parameters for the minimum peak height and number of voxels can be set. clouds that do not fulfil these criteria are considered 'unusable' and do not appear in the final assignment array and catalogue.

A.2 Input parameters

The FW algorithm is implemented within the function `findclumps` in the JCMT Starlink CUPID package. In this function, the emission extraction is regulated by a configuration

file that defines the values of the input parameters. List of parameters that define the extraction [1] is reproduced below,

- `AllowEdge` : If set to a zero value, then clouds are rejected if they touch any edge of the data array. If non-zero, then such clouds are retained.
- `CleanIter`: This gives the number of times to apply the cellular automaton which cleans up the filled clouds.
- `FwhmBeam`: The FWHM of the instrument beam, in pixels. If the deconvolution option is chosen in the `findclumps` function, the cloud widths written to the output catalogue are reduced (in quadrature) by this amount (see below).
- `MaxBad`: The maximum fraction of pixels in a cloud that is allowed to be adjacent to a bad pixel. If the fraction of cloud pixels adjacent to a bad pixel exceeds this value, the cloud is excluded.
- `MinDip`: If the dip between two adjacent peaks is less than this value, then the peaks are considered to be part of the same cloud.
- `MinHeight`: If the peak value in a cloud is less than this value then the cloud is not included in the returned list of clouds.
- `MaxJump`: Defines the extent of the neighbourhood about a local maximum which is checked for higher pixel values. The neighbourhood checked is square or cube with a side equal to twice the supplied value, in pixels.
- `Noise`: Defines the data value below which pixels are considered to be in the noise. No walk will start from a pixel with a data value less than this value.
- `RMS`: The global rms noise level in the data. The default value is the value supplied for parameter rms.
- `VeloRes`: The velocity resolution of the instrument, in channels. The velocity width of each cloud written to the output, the catalogue is reduced (in quadrature) by this amount.

[1] `http://www.starlink.ac.uk/star/docs/sun255.htx/un255ss5.html#Q1-11-37`

A.3 Output catalogue and cloud assignments

The results of the FW cloud extraction in the Starlink suite are presented both as the CAA (a mask that matches the size of the input emission array in which unique cloud identifiers mark the voxels belonging to the individual clouds) and a catalogue that collects certain positional and geometric characteristic of the clouds. Assuming that the extraction is performed on a three-dimensional datacube, where the indices 1 and 2 denote the spatial coordinates, and 3 the spectral axis, the basic FW catalogue will include the following columns,

- `Peak1, Peak2, Peak3` : The position of the cloud peak value on each axis.
- `Peak`: The peak value in the cloud.
- `Cen1, Cen2, Cen3`: The position of the cloud centroid on each axis.
- `Size1, Size2, Size3` : The size of the cloud along each axis (in pixels).
- `Sum`: The total data sum in the cloud (i.e. the sum of the pixel values within the cloud)
- `Volume`: The total number of pixels falling within the cloud.

The size S_i of a cloud in the direction i is measured as the rms deviation of each voxel centre from the cloud centroid C,

$$S_i = \sqrt{\frac{\sum d_i x_i^2}{\sum d_i} - C^2}, \tag{A.1}$$

where

$$C = \frac{\sum d_i x_i}{\sum d_i}. \tag{A.2}$$

The weights d_i are the data values at the voxels minus an estimate of the background value in the cloud (Berry, 2015). If cloud data form a Gaussian distribution, this definition of size coincides with the standard deviation of the distribution.

If the beam of the telescope is known, `findclumps` includes a correction option to remove the instrumental blurring and recover the intrinsic source sizes. When the beam correction is selected, the size defined in A.1 becomes

$$S_{\text{corr}} = \sqrt{s_i^2 - b^2}, \tag{A.3}$$

where b is the size of the telescope beam.

Correcting for the beam also affects the peak value in the cloud. This difference increases as the cloud volume decreases. Assuming the cloud possesses a Gaussian profile and that the sum of the data values within the corrected cloud equals the corresponding sum in the uncorrected cloud, the new peak value becomes

$$\text{peak}_{\text{corr}} = d_{\text{max}} \sqrt{\frac{\text{size1} \cdot \text{size2} \cdot \text{size3}}{\text{size1}_c \cdot \text{size2}_c \cdot \text{size3}_c}}, \tag{A.4}$$

where the subscript c refers to the beam corrected sizes, and d_{max} is the observed peak values. The full FW catalogue published by Rigby et al. (2019) is derived from these quantities after assigning distances, excitation temperatures and masses.

Appendix B

Spectral Clustering for Interstellar Molecular Emission Segmentation

The next sections provide a brief introduction to the theory behind the construction of the Spectral Clustering for Interstellar Molecular Emission Segmentation (SCIMES) method with a focus on the application of abstract graph theoretical concepts to the identification of GMCs in PPV datasets. Cloud recognition through SCIMES thus relies on the transitions in the emission structure in the ISM to define objects and it was shown to provide robust results against changes of the dendrogram-construction parameters, noise realizations and degraded resolution (Colombo et al., 2015a, 2014, 2019). This approach to the segmentation of molecular emission mitigates the problem of over-segmentation of the CO emission caused by high resolution, generates physically oriented cloud catalogues, and has the major advantage of being suitable for application to data sets with wide spatial dynamic ranges (many resolution elements within a single cloud) (Jain et al., 1999; Colombo et al., 2015a).

This appendix is based on the description of the SCIMES algorithm published by Colombo et al. (2015a).

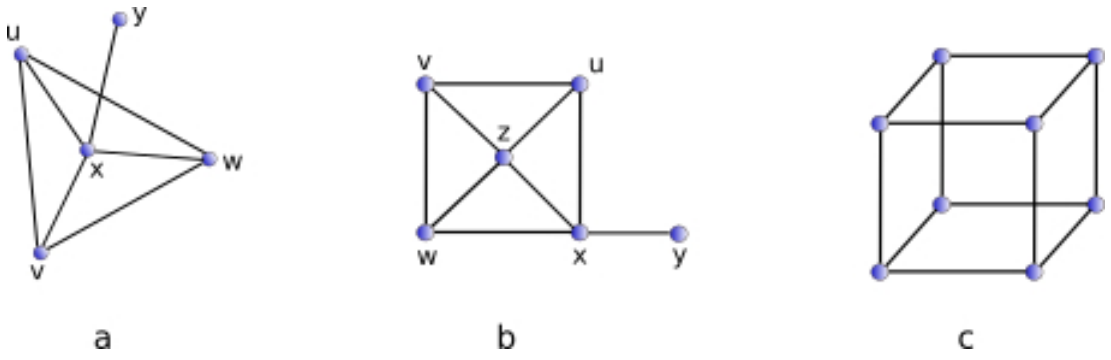


FIGURE B.1: Stick-and-balls representation of two graphs (a and b) with vertex sets (x,y,u,v,w) and (x,y,z,u,v,w) and edge sets (xy, xv, xu ,xw, uv,vw,uw) and (xy, xz, xu, xw, uv, uz, vz, vw, zw) respectively. The terms edge and vertex originate from geometric solids: a cube, for instance, has edges and vertices that represent the graph drawn in panel (c) (West, 2002)

.

B.1 Graphs

This section is a short overview of some general graph theoretical definitions and concepts to establish the terminology used throughout the exposition of the SCIMES algorithm.

A graph G is a triple consisting of a vertex set $V(G)$, an edge set $E(G)$, and a relation that associates with each edge two vertices (not necessarily distinct) called endpoints.

A subgraph of a graph G is a graph H such that $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$ and with the same assignment of endpoints as G.

A graph is drawn by placing each vertex at a point and representing each edge as a curve joining the locations of its endpoints. A graph is called simple when it has no loops or multiple edges (i.e. edges whose endpoints are equal and edges having the same endpoints). A simple graph can be specified by its vertex and set edge sets, considering the latter as a non-ordered set of pairs of vertices. The notation $e = uv$ or $e = vu$ is used to denote the edge e with endpoints u and v (West, 2002). The vertices u and v are adjacent and neighbours. The edge e, and vertices u and v are said to be incident. In a simple graph, the number of edges incident to a vertex constitutes the degree of the vertex.

A path is a simple graph whose vertex set can be ordered so that two vertices are adjacent if and only if they are consecutive in the list. A cycle is a path of edges and vertices in which each vertex is reachable from itself. A simple graph is complete if its vertices are pairwise adjacent. A complete graph is an example of connected graph, a graph in which there is a path between every pair of vertices. Sometimes the name strongly connected is used to refer to a connected graph, while weakly connected is used to denote a graph that includes disconnected parts (not every vertex can be reached through a path starting at any of the other vertices).

B.1.1 Similarity matrix

Let G be a simple graph with vertex set $V(G) = \{v_1, v_2, \ldots, v_n\}$ and edge set $E(G) = \{e_1, e_2, \ldots, e_n\}$. The adjacency matrix $\mathbf{A}(G)$ of G is defined as the $n \times n$ within entries $a_{i,j}$ that correspond to the number of edges in G with endpoints $\{v_i, v_j\}$.

The adjacency matrix thus fully encodes the graph providing a natural representation that is well-suited for computational purposes. The adjacency matrix of a simple graph contains only 0s and 1s. Simple graphs are often used to express relations within a set of entities (see clustering below). The strength/degree of relation between two vertices can be represented as a numerical label associated with each edge. Such a graph is known as a weighted graph. The adjacency matrix of a weighted simple graph can be recast as a similarity (or affinity) matrix. Each entry $s_{i,j}$ of the similarity matrix $\mathbf{S}$ correspond to the weight associated to the edge $\{v_i, v_j\}$. For a weighted graph, the generalised degree of vertex v_i is defined as

$$d_i = \sum_{j=1}^{n} s_{i,j}.$$

The degree matrix, $\mathbf{D}$, of a simple graph is a diagonal matrix that contains the degrees d_i of the vertices v_i on the main diagonal.

B.1.2 Laplacian matrix

The Laplacian matrix $\mathbf{Q}$ of a graph G is the matrix

$$\mathbf{Q} = \mathbf{D} - \mathbf{S},$$

where $\mathbf{D}$ is the degree matrix of G and $\mathbf{S}$ is its similarity matrix. For a weakly connected, simple, weighted graph, the entry of the Laplacian Q_{ij} equals the degree of the vertex v_i when $i = j$ and Q_{ij} is the negative weight of the edge ij when v_i and v_j are adjacent.

The graph Laplacian represents the discrete counterpart of the Laplacian operator ∇^2 (i.e. the multivariable second derivative), applied to a graph. Vertices with a higher degree in a graph (denser nodes on a network) are equivalent to "bumps" in the second derivative of a continuous function, expressing larger changes in the flux density of the gradient flow of the function (Arfken & Weber, 2005).

The list of eigenvalues of Q is called the Laplacian spectrum. The Laplacian spectrum encodes the global properties of the graph it represents. For instance, the number of connected components of a graph corresponds to the multiplicity of the 0 eigenvalue of Q (West, 2002). The Laplacian matrix can be recast as a block diagonal matrix through appropriate permutations so that each connected component of the graph is represented by a block. Since each of these components (subgraphs) is strongly connected, its graph Laplacian has only a single eigenvalue equal zero. Since a graph Laplacian is positive-semidefinite, its second smallest eigenvalue is greater than zero. This eigenvalue is known as the spectral gap. The spectral gap represents the algebraic connectivity of the graph and quantifies how well-connected/dense the graph is (the highest the value, the more connected the graph). The second non-zero eigenvalue eigenvalue is called the Fiedler value. The Fiedler value approximates the minimum number of graph cuts (edge removals) that are needed to partition the graph into two connected components. The components of eigenvector corresponding to the Fiedler value (the Fiedler vector) provide side of the cut each vertex belongs to (spectral graph partitioning).

Often, a symmetric normalized form of the Laplacian is used (Ng et al., 2001):

$$\mathbf{L}_{\text{sym}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{S})\mathbf{D}^{-\frac{1}{2}}, \tag{B.1}$$

since it produces more general eigenvalues, better related to other graph invariants and directly connected to the graph's spectral geometry (Chung, 1997).

B.2 Dendrograms

A dendrogram is a tree diagram that is used to illustrate the hierarchy of structures within a set of data. A dendrogram is defined as a set of two types of structure: the branches and the leaves. The branches are 'subtrees' of the dendrogram. They are, in turn, characterised by multiple substructures: their own branches and leaves. A leaf has no substructure, it is simply a node in the dendrogram. The term trunk is used to refer to a structure that has no parent structure. The 'nested' nature of branches in a dendrogram allows hierarchical structures to be adequately represented. In particular, they can be adopted to provide an abstract representation of the topology of star-forming complexes by encoding the nested spatial arrangement of three-dimensional contours (isosurfaces) at given molecular emission levels in PPV datasets.

In a dendrogram, each point can be intuitively interpreted as defining an isosurface at a fixed emission level. In this context, the leaves of the dendrogram correspond to those isosurfaces with a single local maximum (see Figure 3.1). Such leaves thus form the top of the dendrogram. The branches are represented as vertical segments connecting two leaves, while the horizontal lines mirror the spatial distribution of the emission profile (Figure 3.2). The length of each branch is proportional to the number of contour levels over which the emission properties (such as temperature, intensity) [1] icantly (although the volume of the isosurfaces does change, Rosolowsky et al., 2008).

To discard contamination arising from noise fluctuations, local maxima are determined through a multi-step elimination process. First, each maximum is identified as the voxel with the largest emission value within a box, whose size is determined by significant spatial and spectral resolution elements. Then, the elimination of local maxima proceeds as illustrated in Figure 3.2.

[1] The significant properties are chosen according to a connectivity or similarity criterion that is used to define a GMC as a set of connected voxels.

- A peak is eliminated if its emission is below a set level, `min_val`. The minimum emission level is generally chosen to be a multiple of the root mean square of noise fluctuations(`min_val`= $n\sigma_{rms}$).
- A local maximum is removed if it belongs to an isosurface with a volume smaller than a specified number of voxels (`min_pix`).
- A local maximum is removed if the difference between the peak and the value of the emission at the contour level where it merges with a neighbouring peak is smaller than a threshold value (`min_delta`). The contour profile that contains both peaks is counted as a single local maximum.

The contour level at which two isosurfaces merge is called a merger level. At lower emission levels, all the branches and leaves eventually merge into the trunk of the tree structure. The rules for peak elimination and isosurface mergers defined above force the construction of a dendrogram in which only binary mergers are generated (Rosolowsky et al., 2008).

In SCIMES, the construction of the dendrogram of the molecular emission and the catalogue of the structures it represents rely on the Python dendrogram implementation ASTRODENDRO[2]. This package produces a dendrogram following the criteria specified in the list above once an initial parameterisation is provided. The three input parameters, illustrated in Figure 3.1, specify the emission threshold (`min_val`) below which no structure is considered in the dendrogram (this is usually, a multiple of the data $\sigma_{\rm rms}$); the value (`min_delta`) expressing when a peak is to be counted as an independent leaf (also set to a multiple the observation sensitivity); and the minimum number of pixels (`min_pix`) that must be contained with a leaf (usually, a multiple of the observation beam).

B.3 Dendrogram graph

Dendrograms encode all the information on the topology of molecular emission, however, alone a dendrogram is not enough to precisely identify molecular clouds in a PPV data

[2] `http://www.dendrograms.org`

set. Cloud identification requires a robust mathematical method that uses the properties of the data exclusively to select 'cuts' in the dendrogram's tree structure.

In turn, these partitions in the dendrogram define independent sets in the data that are identified as emission clouds. The first step towards this characterisation of the data set is interpreting the dendrogram as a graph whose vertex set represents the objects on which to apply spectral clustering and consequently induce cuts on the dendrogram.

A vertex set is constructed by considering the leaves (local maxima) in the dendrogram. Any two vertices are then connected by an edge representing the highest level isosurface that contains both leaves. Since all structures (leaves and branches) are connected at the bottom of the dendrogram through the trunk, which represents the union of all the isosurfaces the dendrogram comprises, any vertex is connected to all the others. Graphs associated with dendrograms (or dendrogram graphs for short) are thus complete, simple (no loops since they are meant to represent the relations between pairs of leaves exclusively) and undirected (by definition of edge, the relations between leaves are symmetric).

The edge set associates a structure at a certain hierarchical level to every pair of leaves. This structure defines 'similarity' relations between the leaves. The strength of this association is quantified by assigning weights to the edges. Choosing a good weighting scheme among the many that are possible is crucial in the application of the spectral clustering algorithm (see sections B.4 and B.5.3). This method uses the properties of the similarity matrix (defined in section B.1) alone to find optimal cuts in the graph without providing information a priori on the final cluster assignments. Also, in the context of hierarchical structures, the notions of similarity and distance are usually strictly connected: highly similar objects are likely to be found within a short distance from one another.

B.4 Similarity matrix

An affinity or similarity relation applied to a dendrogram graph (see section B.3) defines its similarity matrix (see section B.1). The SCIMES method implements two weighting schemes that focus on the luminosity and volume of the structures identified by the

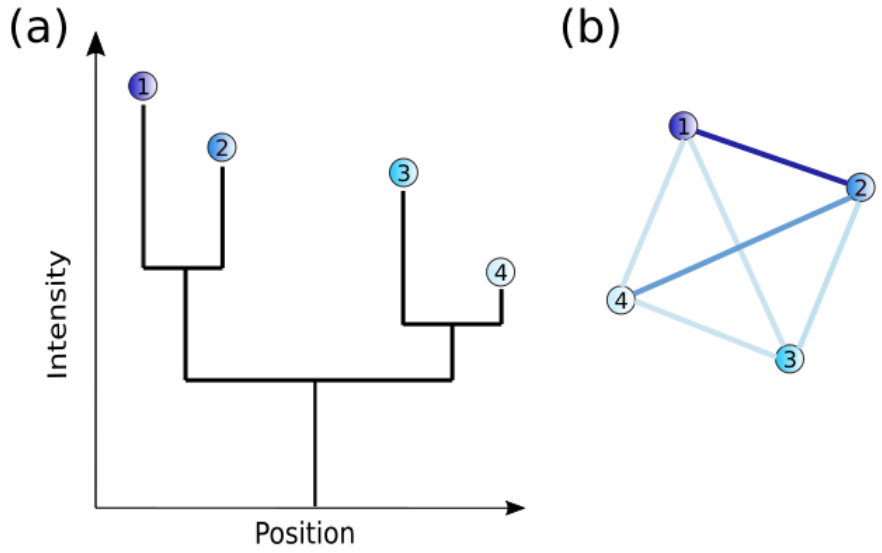


FIGURE B.2: A dendrogram (a) and the graph associated to it (b). The color of the edges encodes the strength of the connection between the leaves (the darker the color, the higher the weight). The weight assigned to an edge between leaves reflect the degree of hierarchical intensity level, e.g. leaves 1 and 2 exhibit a higher hierarchical connection than leaves 2 and 3. Since all leaves are connected trough the trunk, the dendrogram defines a fully connected (complete) graph.

dendrogram. This section explains the criteria and measurements that SCIMES uses to construct the similarity matrices.

B.4.1 Luminosity

Consider the molecular emission in a PPV data set (coordinates x, y, v), and let T_i be the brightness temperature at a voxel at position x_i, y_i and v_i and the size of the voxels be $\delta x \times \delta y \times \delta v$. The flux within an isosurface is then given by the sum of the emissions of the voxels it comprises (Rosolowsky & Leroy, 2006). The flux is defined as

$$F = \sum_i T_i \delta x \delta y \delta v.$$

Assuming a physical distance, d, the luminosity of the isosurface can be derived as

$$L = F d^2. \tag{B.2}$$

B.4.2 Volume

Consider the projection onto the $x-y$ plane of a structure in PPV space. Principal component analysis (Jolliffe & Cadima, 2016) allows us to find the major and minor axes of the projection. Introducing a rotation that aligns the major axis of the projected structure with the x-axis of the coordinate systems and the minor axis with the y-axis, makes it possible to derive the root mean squared sizes of the structure using the intensity weighted second moments along the x and y (Rosolowsky et al., 2008; Colombo et al., 2015a) as

$$\sigma_{\rm maj} = \sqrt{\frac{\sum_i (T_i x_i - \bar{x})}{\sum_i T_i}}, \tag{B.3}$$

and

$$\sigma_{\rm min} = \sqrt{\frac{\sum_i (T_i y_i - \bar{y})}{\sum_i T_i}}, \tag{B.4}$$

where the notation introduced in sub-section B.4.1 was used and the symbol ' $^-$ ' denotes the mean value along an axis.

Similarly, the velocity dispersion in the spectral direction v is

$$\sigma_{\rm v} = \sqrt{\frac{\sum_i (T_i v_i - \bar{v})}{\sum_i T_i}}, \tag{B.5}$$

From B.3 and B.4 the root mean squared size of the structure can be obtained

$$\sigma_{\rm r} = \sqrt{\sigma_{\rm maj}\sigma_{\rm min}}. \tag{B.6}$$

The volume of a spherical cloud with the same root mean square size can be calculated with the radius $R = \eta\sigma_{\rm r}$ with $\eta = 1.91$ (Solomon et al., 1987; Rosolowsky & Leroy, 2006). Finally, using the second similarity criterion and the velocity dispersion (Rosolowsky & Leroy, 2006), the volume of the isosurface is given by

$$V = \pi R^2 \sigma_{\rm v}. \tag{B.7}$$

B.4.3 On distances

Similarity relations based on luminosity and volume are general criteria and they involve information on the distance of the sources considered. In Galactic surveys of molecular emission, however, precise estimations of distances are rarely available. In the absence of known distances, both the volume and luminosity criterium need to be modified. For luminosities, equation B.2 is simply set to $L = F$. A flux criterion is used instead, for which F has units of K km s^{-1}. While the volume criterion (equation B.7) retains its form, but is interpreted as measured in arcsec^2 km s^{-1} . A comparison of these similarity criteria on image segmentation and cloud identification through SCIMES is presented in (Colombo et al., 2015a) for the Orion Monoceros region. By default SCIMES considers the "volume" and "luminosity" matrices. However, the user defined affinity matrices can also be provided to produce a segmentation based on some specific property of the ISM [3]. Such matrices must be ordered following the indexing of the the dendrogram leaves. Multiple similarity matrices can be provided at the same time. In this case, SCIMES will aggregate them and produce a segmentation based on all of the given criteria.

B.4.4 Weighting schemes

The weight of an edge reflects the properties of the highest emission level at which adjacent leaves merge. This merger level corresponds to a brightness temperature isosurface. By definition of emission dendrogram (section B.2), the properties of an isosurface are largely unchanged within a branch of the dendrogram and they usually depend on the contour level continuously. Continuity is lost at the merging points of branches. The merger surface, in fact, contains more emission than any of its individual branches. In general, the size of the isosurfaces is inversely proportional to their hierarchical level. Thus, the weight of an edge will also be inversely proportional to the chosen properties of the emission of its corresponding merger surface. In the case of our similarity criteria, smaller volumes and lower luminosities/fluxes have heavier weights. Formally, let i and j

[3] SCIMES works best with monotonic and block diagonal matrices. Non-monotonic and strictly continuous similarity criteria could produce errors in the clustering process and the resulting segmentation (Colombo et al., 2015a)

be two vertices of the dendrogram graph, and L_{ij} and V_{ij} the luminosity and the volume of the isosurface corresponding to the edge ij. Then the weights assigned to ij are

$$W_{ij}^{L} = \frac{1}{L_{\mathrm{ij}}},$$

and

$$W_{ij}^{V} = \frac{1}{V_{\mathrm{ij}}}.$$

By definition of edge in the dendrogram graph, the similarity matrix is symmetric.

B.4.5 Rescaling

The strength of a similarity relation on the dendrogram graph defines the local neighbourhood relations of each leaf. The stronger the relation, the closer the neighbour. In order for the similarity matrix to enhance this feature, it is often smoothed with a kernel function. Gaussian kernels are often used in this practice:

$$s_{ij} = \exp\left(-\frac{w_{ij}^2}{2\sigma_s^2}\right), \tag{B.8}$$

where s_{ij} is the rescaled version of the weight w_{ij} on the edge ij. The smoothing parameter σ_s controls the scaling of the size of the local neighborhood of the leaves i and j. In other words, σ_s determines how quickly the similarity between two leaves declines with distance. The value of σ_s affects the resulting clustering partition of the dendrogram: choosing too small a value for σ_s produces a similarity matrix where only the weights of directly neighboring leaves are significant, on the other hand, a large σ_s blends neighborhoods and results in under-clustering (Colombo et al., 2015a).

Fischer & Poland (2004) show that it is possible to estimate an appropriate value of σ_s by constructing a "similarity histogram" and considering its modes. Such a histogram is simply the result of binning the weights of the dendrogram graph. If the leaves of

the dendrogram graph can be collected in clusters according to the chosen similarity criteria, then the histogram has multiple modes. In particular, the first mode occurs at the average similarity (weight) within clusters, while the other nodes represent the similarities between clusters. A smoothing value that lies between the first two nodes, is then expected to strengthen the weights between intercluster leaves while weakening those of intracluster leaves. For instance, choosing the median value between the first two modes ensures both under-clustering and over-clustering the data set is avoided.

As a rule of thumb for spectral clustering, a good input similarity matrix has a block-diagonal form (obtained after multiple permutation of its rows and columns) with each block having similar entries on its boundary (Fischer & Poland, 2005; von Luxburg, 2007; Colombo et al., 2015a).

B.4.6 Matrix aggregation

Shi & Malik (2000) show that different similarity criteria can be combined into a single similarity matrix. This operation is known as matrix aggregation and is applied by the authors to a color image segmentation problem. Following Shi's method, the SCIMES algorithm considers the volume and luminosity matrices after rescaling with the appropriate kernel, and 'aggregate' them through element-wise multiplication. The resulting product and the volume and luminosity matrices serve as the main input for the spectral clustering algorithm (Shi & Malik, 2000).

B.4.7 Observations

By default, luminosity and volume are adopted as clustering criteria in SCIMES. Both luminosity and volume are good indicators of similarity in emission structures. They describe physical properties (emissivity, velocity, and morphology) of molecular emission structures. Thus, they allow for the identification structure and sub-structure in both spatial and spectral directions through the differences in emission. In addition, volume and luminosity increase monotonically (and discontinuously) as the level of dendrogram hierarchy decreases (isosurfaces increase in volume and consequently their flux rises). Discontinuities in luminosity and volume are especially apparent when two surfaces with

similar characteristics merge. The hierarchy levels described by the dendrogram itself are monotonic in terms of the number of isolate isosurfaces (the higher the level, the greater their number). These features make the volume and luminosity criteria result in well-behaved, block diagonal (after row and column permutation) similarity matrices. The analysis of this form of the similarity matrix makes it possible to estimate the approximate number of final clusters (see section B.5).

B.5 The SCIMES algorithm

Equipped with a re-scaled similarity matrix that embodies the strength of the relations between the leaves of the dendrogram graph, a clustering method to partition the dendrogram can now be introduced. The intuition of clustering is to partition a set of data points into subsets whose elements have a comparable degree of similarity according to their similarity relations. In the case of a dendrogram graph, a partition in which the edges between the leaves in the same cluster have higher weights than the edges that connect them to leaves in other clusters is searched for. The cuts in the dendrogram defined by these clusters of leaves are then identified in the molecular emission data as individual, independent objects, the giant molecular clouds.

Spectral clustering uses the eigenvectors of Laplacian derived from the similarity matrix (see section B.1) to translate the clustering problem from the space of $n \times n$ matrices to a lower- dimensional metric space (**spectral embedding**). In this new space, the initial similarity relations are identified with Euclidean distances. A standard k-means clustering algorithm can thus be applied to these new sets of data points. The resulting clustering scheme provides an optimal partition of the dendrogram graph based on the number of clusters provided as input. Spectral clustering is particularly efficient on complete, weighted, undirected, and simple graphs (von Luxburg, 2007).

B.5.1 Algorithm (spectral clustering)

Consider a dendrogram graph G with vertex set $V(G)$, and its similarity matrix $\mathbf{S}$ (the aggregate matrix of the volume and luminosity criteria),

- Input: a similarity matrix $\mathbf{S}$ $(n \times n)$ and an estimated number of clusters k.

  1. Construct the degree matrix $\mathbf{D}$ and normalized symmetric Laplacian $\mathbf{L}$.
  2. **Spectral embedding**: compute the set of the k largest eigenvalues[4] of $\mathbf{L}$, consider their corresponding eigenvectors $(u_1, u_2, \ldots, u_k)$.
  3. Construct the **eigenvector matrix**: a matrix $\mathbf{U} \in \mathbb{R}^{n \times k}$, $k < n$, whose columns are the eigenvectors of $\mathbf{L}$.
  4. Consider the set of n vectors $(y_i)_{i=1,2,\ldots,n} \in \mathbb{R}^k$ that correspond to the rows of $\mathbf{U}$.
  5. Apply **k-means algorithm** to collect the points $(y_1, y_2, \ldots, y_n)$ into the clusters $C_1, C_2, \ldots, C_k$.

- Output: $A_1, A_2, \ldots A_k \subset V(G)$, such that $\bigcup_{i=1}^{k} = V(G)$ and $A_i \cap A_j = \emptyset$ for any i and j. If $y_i \in C_l$ then $v_i \in A_l$.

The success of spectral clustering is greatly due to its absence of assumptions on the shape of the clusters it generates (as opposed to k-means, for which the resulting clusters are always convex hulls, see section B.5.4). Spectral clustering can thus be applied to very general problems and complex distribution of data points. In addition, spectral clustering is efficient on very large data sets as long as the input similarity matrix is sparse (von Luxburg, 2007). For a given similarity matrix, the algorithm solves a linear problem, without the risk of getting stuck in local minima or requiring several runs with different initializations.

B.5.2 The silhouette coefficient

In order to apply spectral clustering to the Laplacian of the dendrogram graph, the number of clusters into which the algorithm is to arrange the data must be provided. Such an input parameter is common to many clustering algorithms. Different methods have been devised to estimate its best possible value from theoretical and statistical analysis of the data (Tibshirani et al., 2001; Still & Bialek, 2004). In the particular case of spectral clustering, the number of clusters k can be either evaluated from the analysis the spectrum of eigenvectors of the Laplacian (Zelnik-Manor & Perona, 2004) or by

[4] Eigenvalues with multiplicity greater than 1 are all included in the set.

assessing the quality of clustering through special measures. The latter method is based on measuring the ratio of the similarities (weights) of the intra- and intercluster data points. Such a measure can be directly evaluated from the similarity matrix (Rosseeuw, 1987).

Consider an object/point i in a set objects with a similarity relation, the silhouette coefficient of i is defined as

$$\mathrm{sil}(i) = \frac{b(i) - a(i)}{\max(a(i) - b(i))}, \tag{B.9}$$

where $a(i)$ is the average similarity between the point i and all other points in the same cluster and $b(i)$ is the average similarity between i and all the points in the next nearest cluster. For a point i, $\mathrm{sil}(i) \in [-1, 1]$. The value of $\mathrm{sil}(i)$ contains information on the nature of the clustering, in particular,

- $\mathrm{sil}(i) = -1$ for incorrect clustering,
- $\mathrm{sil}(i) = 0$ for overlapping clusters,
- $\mathrm{sil}(i) = 1$ for high intracluster similarity and low intercluster similarity.

Thus, increasingly positive values of $\mathrm{sil}(i) = 1$ indicate denser and better-separated clusters. The average of value $\mathrm{sil}(i)$ over all data points provides a measure of how well the data have been partitioned. Since the average silhouette depends on k in a non-monotonic way, optimization techniques such as genetic algorithms are usually employed to determine the number of clusters that maximize the silhouette (Lleti et al., 2004).

In SCIMES, an initial value for k is thus guessed after rescaling the similarity matrix via an appropriate kernel function (see subsection B.4.5). A suitable σ_s enhances the similarity relations and the blocks with the heaviest weight can be isolated as related to the final clustering configuration [5] An iterative optimisation is then run from this initial k to maximise the average silhouette. In SCIMES, silhouette optimisation is handled by the Python SCIKIT–LEARN package[6].

[5] This operation is similar to using the Fiedler vector (the eigenvector of the Laplacian that corresponds to the second smallest eigenvalue) to determine the algebraic connectivity of a graph (Fiedler, 1973).

[6] `http://scikit-learn.org/stable/modules/clustering`

B.5.3 Spectral embedding

The core of the spectral clustering algorithm is spectral embedding, a transformation that performs a dimension reduction by mapping the data in the dendrogram graph into a 'cluster' vector space (R^k). This operation translates the description of similarity between graph vertices into Euclidean distances in the new metric space (see Figure B.3). In $\mathbb{R}^k$, points corresponding to highly similar leaves are grouped together, making clustering patterns based on similarity easily identifiable. Spectral embedding relies on properties of the graph Laplacian. The elements of the eigenvectors corresponding to the first k largest eigenvalues provide a k-dimensional description of the block structure of the Laplacian and the k components of the graph with the highest algebraic connectivity.

B.5.4 k-means algorithm

In a vector space with Euclidean metric, the sets of data points are easily grouped together with common clustering algorithms. SCIMES uses k-means (MacQueen, MacQueen) in R^k to find the configurations of k clusters of the data points that maximise the intracluster distance and minimise the intercluster distance. This algorithm is known for its fast convergence (Arthur & Vassilvitskii, Arthur & Vassilvitskii).

Given an estimated number of clusters k, the algorithm

- selects k means or centroids randomly,
- associates each data point to the nearest centroid (Euclidean distance)
- calculates the position of the centroids of these clusters,
- iterates the last two steps until convergence is reached (the new centroids are exactly in the positions of the ones found before).

This model considers spherical clusters that are separable so that the centroids converge towards a clusters' center upon iteration. For the assignment of a point to the center of the nearest cluster, clusters are expected to be of similar size. The result of the k-means algorithm can be interpreted as the Voronoid cells of the cluster centroids, with data

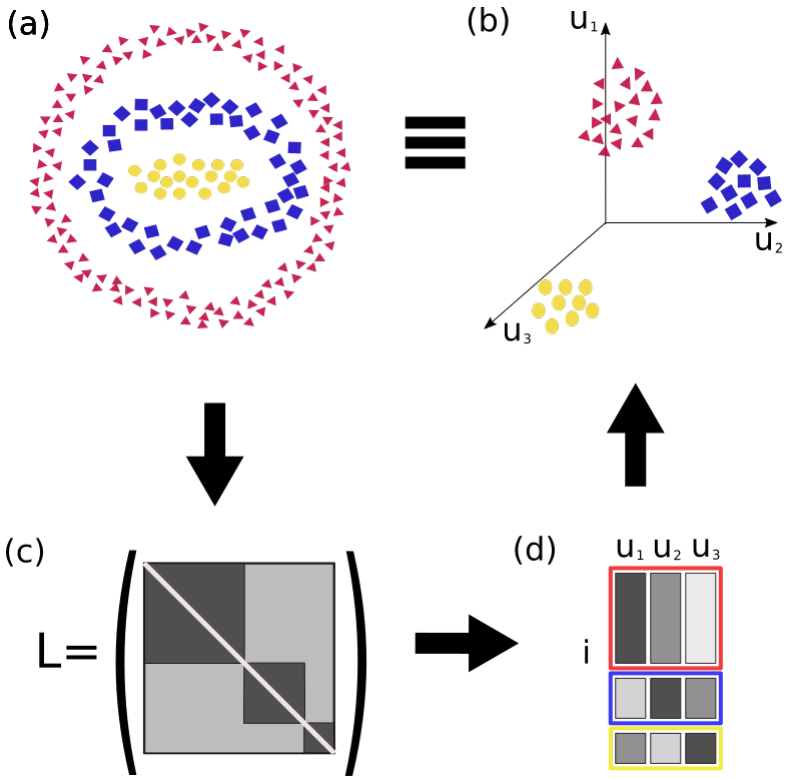


FIGURE B.3: Representation of spectral embedding. The initial distribution of objects is depicted in (a), where highly similar objects have the same color and shape. Panel (b) shows the graph Laplacian for the distribution in (a). Choosing a good similarity criterion for clustering, produces, after an index permutation, a block diagonal graph Laplacian. In the Laplacian, pairs of objects of highly similar objects are colored black, while grey is used for lower similarity objects. The degrees of the graph vertices are located on the main diagonal. The eigenvectors corresponding to the largest k eigenvalues are arranged in a matrix in (c). The estimated number of clusters (silhouette maximization) defines the number eigenvectors considered and the dimension of the 'clustering' space ($\mathbb{R}^k$). Here $k = 3$ is considered and every vertex/leaf v_i of the initial dendrogram graph is represented as a point in $\mathbb{R}^3$ with coordinates $(u_1()i), u_2(i), u_3(i))$ as shown in (d). In the embedded clustering space, the initial distribution is remapped to well-separated collections of objects. This new distribution can be clustered using k-means and Euclidean distances. Picture and explanation after (Colombo et al., 2015a).

points being separated halfway between clusters' centroids (Aurenhammer, 1991). This tessellation may lead to non-optimal clustering (see Figure B.4 for an example produced with ELKI[7]) with points of a cluster that have no nearest neighbours belonging to that cluster. In PPV space, such leaves are collected into sparse clusters without any neighbours between constituent objects. These leaves are eliminated from the final labelling of clusters.

[7] https://elki-project.github.io/tutorial/same-size_k_means

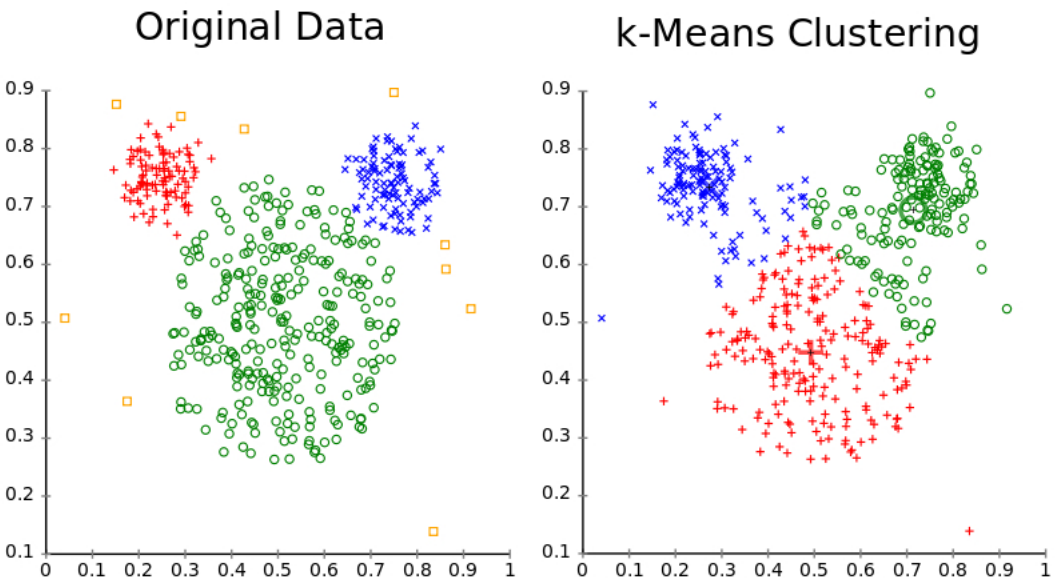


FIGURE B.4: Clustering of an artificial dataset ('the mouse') with the k-means algorithm. As k-means tend to produce clusters of similar sizes, some data points have no nearest neighbours belonging to that cluster. Data visualisation generated with ELKI.

B.6 Final cloud identification

The final clusters selected through spectral clustering correspond to branches of the dendro- gram that contain only leaves in a single cluster. These branches make up a partition of the dendrogram. Similar leaves that do not form isolated compact clusters in PPV space are collected in sparse clusters. These sets of objects (with no neighbouring emission peaks) are considered as noise artefacts therefore removed from the final labelling of the clusters. The remaining clusters are emission structures that were already considered by the original dendrogram algorithm. They represent the relevant independent molecular clouds embedded in the emission. Since rescaling the similarity matrix enhances the clustering of leaves above a threshold value of luminosity and volume, the final selection of clouds presents similar properties (in terms of luminosity and volume), but with clouds located at different hierarchical levels of the emission structure (Colombo et al., 2015a).

B.7 Cluster and leaf assignments

The main output of the SCIMES algorithm, consists of a list of dendrogram indices corresponding to the relevant structures within the emission dendrogram. Recall that these structures are already encoded within the dendrogram and their hierarchy can be accessed through the ASTRODENDRO class methods[8]. In addition, the package Astrodendro collects the physical and geometric properties and in the dendrogram PPV catalog. In addition, the get_mask method of Astrodendro is called to construct an assignment cube of the clouds. Pixels within each cloud are uniquely labelled with a number corresponding to the index of the structure in the dendrogram. The method automatically generates cubes for identified cluster, leaf , and trunk structures which are saved as fits images.

B.8 Cloud catalogue

The properties of the structures resulting from the SCIMES segmentation are collected in a catalogue constructed through the ASTRODENDRO PPV statitics. The entries in the catalogue are listed below as they are defined in the ASTRODENDRO documentation website[9].

`major_sigma` : Major axis of the projection onto the position-position plane, computed from the intensity weighted second moment in direction of greatest elongation in the PP plane.

`minor_sigma` : Minor axis of the projection onto the position-position plane, computed from the intensity weighted second moment in direction of greatest elongation in the PP plane.

`area_ellipse` : The area of the ellipse defined by the second moments, where the semi-major and semi-minor axes used are the half-width at half-maximum derived from the moments.

`area_exact` : The exact area of the structure on the sky.

[8] `http://www.dendrograms.org`
[9] `http://www.dendrograms.org`

`radius` : Geometric mean of `major_sigma` and `minor_sigma` (in pixels).

`radius_arcsec` : The `radius` converted to arcsec.

`position_angle` :The position angle between the maximum and minimum sky coordinate in degrees (counter-clockwise from the positive x axis. Notice that this positive x axis in pixel coordinates corresponds the the negative x axis in conventional astronomy images).

`x_cen` : The mean position of the structure in the x direction.

`y_cen` : The mean position of the structure in the y direction.

`v_cen` : The mean velocity of the structure.

`v_rms` : Intensity-weighted second moment of velocity.

`flux` The integrated flux of the structure, in Jy (note that this does not include any kind of background subtraction, and is just a plain sum of the values in the structure, converted to Jy).

`sig_kms` : The velocity dispersion calculated as the product between `v_rms` and the size of the velocity channels.

`volume` : The approximate volume of the cloud estimated from `area_ellipse` and `sig_kms`.

Appendix C

Analysis of turbulence

C.1 Preliminaries

This appendix explains a general method to link the power of a field defined on a three-dimensional space to the power of its two-dimensional projection, obtained by averaging along one coordinate axis.

Consider a physical field $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ defined over a cubic region of side L. The spatial average of $\langle \mathbf{F} \rangle$ of $\mathbf{F}$ over $\Omega \in \mathbb{R}^3$ is then defined as

$$\langle \mathbf{F} \rangle = \frac{\int_\Omega F(\mathbf{x})\, \Omega}{\int_\Omega d\Omega}. \tag{C.1}$$

The variance $\sigma^2_{\mathbf{F}}$ is

$$\sigma^2_{\mathbf{F}} = \langle \mathbf{F}^2 \rangle - \langle \mathbf{F} \rangle^2. \tag{C.2}$$

Introduce the density field $\rho : \mathbb{R}^3 \to \mathbb{R}$ and the velocity field $v : \mathbb{R}^3 \to R^3$, both defined over a cubic region $V \in \mathbb{R}^3$ with side L. Define the ρ^q-weighted velocity dispersion, σ^2_q, on the volume V as

$$\sigma^2_q = \frac{\int_V \rho^q v^2\, dV}{\int_V \rho^q\, dV} = \frac{\langle \rho^q v^2 \rangle}{\langle \rho^q \rangle}, \tag{C.3}$$

where the last equality is obtained by multiplying and dividing by $1/V$ and using the definition of spatial average.

In general $\langle \rho v^2 \rangle / \langle \rho^2 \rangle \neq \langle \rho v^2 \rangle / \langle \rho \rangle$. Equality holds only for a uniform density field or when the density and velocity fields are not statistically correlated. Since none of these conditions is usually satisfied in the ISM, statistical correction factors are required. A method that will help determine the correlation between the density and velocity field will now be discussed. This method provides an estimate of the velocity dispersion weighted by various powers of ρ. Introducing the notation $\rho_0 = \langle \rho \rangle$ and $\xi = \rho/\rho_0$, equation C.3 becomes

$$\sigma_q^2 = \frac{\frac{1}{V}\int_V \xi^q v^2 \, dV}{\frac{1}{V}\int_V \rho^q \, dV} = \frac{\langle \rho^q v^2 \rangle}{\langle \xi^q \rangle}. \tag{C.4}$$

In terms of the probability distribution functions $P_v(v)$ and $P_\xi(\xi)$ of v and ξ, the volume integrals in C.4 can be recast as

$$\sigma_q^2 = \frac{\int_0^\infty \int_{-\infty}^\infty P_\xi(\xi) P_v(v) \xi^q v^2 \, d\xi dv}{\int_0^\infty \int_{-\infty}^\infty P_\xi(\xi) P_v(v) \xi^q \, d\xi dv} \tag{C.5}$$

If velocity and density are correlated, $P_v(v)$ can be cast as an implicit function of ξ, the density-dependent velocity dispersion is defined as

$$\sigma_v^2(\xi) = \int_{-\infty}^\infty P_v(v) v^2 \, dv, \tag{C.6}$$

and equation C.5 can be recast as

$$\sigma_q^2 = \frac{\int_0^\infty P_\xi(\xi) \xi^q \sigma_v^2(\xi) \, d\xi}{\int_0^\infty P_\xi(\xi) \xi^q \, d\xi}. \tag{C.7}$$

Assume that

$$\sigma_v^2(\xi) = h(\xi) \sigma_{00}^2, \tag{C.8}$$

where σ_{00}^2 is a constant and

$$h(\xi) = \xi^{-\epsilon}, \tag{C.9}$$

with ϵ being a small positive constant that makes densities are inversely proportional to velocity dispersion.

Substituting into equation C.6 gives

$$\rho_q^2 = \frac{\int_0^\infty P_\xi(\xi)\xi^{q-\epsilon}\sigma_{00}^2 \, d\xi}{\int_0^\infty P_\xi(\xi)\xi^q \, d\xi} = \frac{\langle \xi^{q-\epsilon}\rangle \sigma_0^2}{\langle \xi^q \rangle \langle \xi^{-\epsilon}\rangle}, \tag{C.10}$$

with

$$\sigma_0^2 = \sigma_{00}^2 \langle \xi^{-\epsilon}\rangle \tag{C.11}$$

being the non-weighted ($q = 0$) velocity dispersion.

If velocity dispersion and density are not statistically correlated, $\epsilon = 0$, equation C.10 yields

$$\sigma_q^2 = \sigma_0^2, \quad \forall \; q. \tag{C.12}$$

Combining equation C.3 and C.10 (without their normalizing factors), it can be seen that for all q's, the moments $\langle \xi^q v^2\rangle(q)$ are linked to $\langle \xi^q \rangle(q)$ through a scaling factor and a translation

$$\langle \xi^q v^2 \rangle = \frac{\langle \xi^{q-\epsilon}\rangle \sigma_0^2}{\langle \xi^{-\epsilon}\rangle}. \tag{C.13}$$

Thus, equation C.13 can be used to convert between velocity dispersions weighted by different powers of ρ. However, $\langle \xi^q \rangle(q)$ cannot be obtained directly from observations. To obviate the lack of observational quantities, an analytical form of $P_\xi(\xi)$ can be considered. Under the assumption of isothermal turbulence, a lognormal probability density function can be chosen (Vàzquez-Semadeni, 1994; Padoan et al., 1997; Federrath et al., 2008b)

$$\langle \xi^q \rangle = \exp\left[q\langle ln(\xi)\rangle + \frac{1}{2}q^2\sigma^2_{ln(\xi)}\right]. \tag{C.14}$$

Normalising the field ξ ($\langle\xi\rangle = 1 \Rightarrow \langle ln(\xi)\rangle = -1\frac{1}{2}\sigma^2_{ln(\xi)}$) and remembering that $\sigma^2_{ln(\xi)} = ln(1+\sigma^2_\xi) = ln(\langle\xi^2\rangle)$, gives

$$\langle \xi^q \rangle = \exp\left[\frac{1}{2}\sigma^2_{ln(\xi)}(q^2-q)\right] = \langle\xi^2\rangle^{\frac{1}{2}(q^2-q)}. \tag{C.15}$$

With this result, equation C.13 becomes

$$\langle \xi^q v^2 \rangle = \langle\xi^2\rangle^{\frac{1}{2}(q^2-q-2q\epsilon)}\sigma_0^2, \tag{C.16}$$

by which equation C.7 can be re-written as

$$\sigma_q^2 = \frac{\langle\xi^q v^2\rangle}{\xi^q} = \langle\xi^2\rangle^{-q\epsilon}\sigma_0^2. \tag{C.17}$$

Finally, the ratio g_{mn} of the velocity dispersions can be defined

$$g_{mn} = \frac{\sigma_m^2}{\sigma_n^2} = \frac{\langle\rho^m v^2\rangle/\langle\rho^m\rangle}{\langle\rho^n v^2\rangle/\langle\rho^n\rangle} = \langle\xi^2\rangle^{(n-m)\epsilon}. \tag{C.18}$$

This expression provides a relation for the conversion between different velocity dispersions weighted by powers of ρ,

$$g_{21} = \frac{\sigma_2^2}{\sigma_1^2} = \frac{\langle\rho^2 v^2\rangle/\langle\rho^2\rangle}{\langle\rho v^2\rangle/\langle\rho\rangle} = \langle\xi^2\rangle^{-\epsilon}. \tag{C.19}$$

Finally, the decomposition of a field into its solenoidal and compressive components is discussed. Let $\mathbf{F} : V \to \mathbb{R}$ be a C^2 vector field defined on a bounded domain $V \in \mathbb{R}^3$ enclosed by the surface S. According to the Fundamental Theorem of Vector Calculus (Helmholtz Decomposition Theorem, Helmholtz (1858)), $\mathbf{F}$, can be decomposed into the sum

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}_\perp(\mathbf{x}) + \mathbf{F}_\parallel(\mathbf{x}), \tag{C.20}$$

where $\mathbf{F}_\perp$ is a purely solenoidal (divergence-free, incompressible, or transversal) component

$$\nabla \cdot \mathbf{F}_\perp = 0 \tag{C.21}$$

given by

$$\mathbf{F}_\perp = \nabla \times \mathbf{A},$$

with

$$\mathbf{A}(\mathbf{r}) = \frac{1}{4\pi} \int_V \frac{\nabla' \times \mathbf{F}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} dV' - \frac{1}{4\pi} \oint_S \hat{\mathbf{n}}' \times \frac{\mathbf{F}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} dS'$$

and $\mathbf{F}_\parallel$ a purely compressible (curl-free, irrotational, conservative, or longitudinal)

$$\nabla \times \mathbf{F}_\parallel = 0 \tag{C.22}$$

given by

$$\mathbf{F}_\parallel = -\nabla \Phi,$$

where

$$\Phi(\mathbf{r}) = \frac{1}{4\pi} \int_V \frac{\nabla' \cdot \mathbf{F}(\mathbf{r}')}{|\ \mathbf{r} - \mathbf{r}'\ |} dV' - \frac{1}{4\pi} \oint_S \hat{\mathbf{n}}' \cdot \frac{\mathbf{F}(\mathbf{r}')}{|\ \mathbf{r} - \mathbf{r}'\ |} dS'$$

and ∇' is the nabla operator with respect to $\mathbf{r}'$.

The decomposition introduced in equation C.20 is unique, up to an addictive (vector) constant. Intuitively, one can add linear terms to Φ and $\mathbf{A}$ that contribute to $\mathbf{F}$ in the form of vector constants, e.g.

$$\Phi \to \Phi + z,$$

gives

$$\nabla\Phi \to \nabla\Phi + \mathbf{e}_z,$$

and

$$\mathbf{A} \to \mathbf{A} + \frac{1}{2}(y\mathbf{e}_x - x\mathbf{e}_y)$$

which yields

$$\nabla \times \mathbf{A} \to \nabla \times \mathbf{A} - \mathbf{e}_z.$$

The vector constants in $\mathbf{F}_\perp$ and $\mathbf{F}_\parallel$ then cancel out in the decomposition C.20. The field $\mathbf{F}$ could also possess a component of the form

$$\mathbf{F}_L = \nabla\phi, \tag{C.23}$$

where ϕ is a scalar harmonic field

$$\nabla^2\phi = 0. \tag{C.24}$$

The Laplacian equation C.24 implies that $\mathbf{F}_L$ is divergence-free. In addition, $\mathbf{F}_L$ is curl-free since it is defined as the gradient of a scalar field. Since ϕ is a harmonic field, the mean value theorem holds: for any $\mathbf{x}$ in the domain of $\phi(\mathbf{x})$, the average value of ϕ of the surface of a ball of arbitrary radius centred at $\mathbf{x}$ equals $\phi(\mathbf{x})$. It follows that Φ attains no local extrema within the boundary of its domain. Thus, the boundary conditions of Φ decide its properties and $\mathbf{F}_L$ $(= \nabla\Phi)$ represents domain-wide smooth gradients, which are not accounted for by $\mathbf{F}_\perp$ and $\mathbf{F}_\parallel$. For fields with periodic boundary conditions, the choice of boundary and the absence of local extrema guarantee that Φ_L is constant and

thus $\mathbf{F}_L = 0$. Thus providing a unique Helmholtz decomposition Brunt & Federrath (2014).

For the decomposition of real physical fields in the ISM (such as the momentum density, see below), it is often desirable to consider isolated clouds in which the field decays smoothly to zero at the surface of the cloud. For these clouds, the boundary conditions do not become problematic. The best candidates for the decomposition of density fields are isolated molecular clouds since it is more challenging to ensure the absence of large-scale gradients in the more extensively distributed atomic gas. In particular, a density-weighted velocity field (see below) is continuous as it transitions from molecular to atomic gas [1].

C.2 General Method

This section presents a general method to link the power of a field defined on a three-dimensional space to the power of its two-dimensional projection, obtained by averaging along one coordinate axis.

Let's start with $\mathbf{F}$ defined over the cubic volume V. The Fourier series of $\mathbf{F}(\mathbf{r})$ over the interval $[-L/2, L/2]$ is given by

$$\tilde{\mathbf{F}}(\mathbf{k}) = \int_{-L/2}^{L/2}\int_{-L/2}^{L/2}\int_{-L/2}^{L/2} \mathbf{F}(\mathbf{r}) e^{-\frac{2\pi i \mathbf{k}\cdot\mathbf{r}}{L}} d\mathbf{r}, \tag{C.25}$$

where is $\mathbf{r}$ the position vector (x, y, z) and $\mathbf{k} = (k_x, k_y, k_z) \in \mathbb{Z}^3$ is the vector of spatial frequencies.

The inverse transform $\tilde{\mathbf{F}}$ of $\mathbf{F}$ can thus be written as

$$\mathbf{F}(\mathbf{r}) = \frac{1}{L^3} \sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \sum_{k_z=-\infty}^{\infty} \tilde{\mathbf{F}}(\mathbf{k}) e^{\frac{2\pi i \mathbf{k}\cdot\mathbf{r}}{L}}. \tag{C.26}$$

Now consider the projection of $\mathbf{F}$ onto the xy-plane, $\mathbf{F}_p : \mathbb{R}^2 \to \mathbb{R}$ constructed through the average of $\mathbf{F}$ along the z-direction:

[1] The restriction to the molecular component is a limitation of modelling the ISM as a single fluid. A full description of the ISM is also challenged by the accessibility of observable regions using trace molecules

$$\mathbf{F}_p(x,y) = \frac{1}{L}\int_{-L/2}^{L/2} \mathbf{F}(x,y,z)dz. \tag{C.27}$$

The transform and inverse transform of $\mathbf{F}_p$ are

$$\tilde{\mathbf{F}}(\mathbf{k_2}) = \int_{-L/2}^{L/2}\int_{-L/2}^{L/2} \mathbf{F}_p(\mathbf{r_2}) e^{-\frac{2\pi i \mathbf{k_2}\cdot\mathbf{r_2}}{L}} d\mathbf{r_2} \tag{C.28}$$

and

$$\mathbf{F}_p(\mathbf{r_2}) = \frac{1}{L^2}\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty} \tilde{\mathbf{F}}_p(\mathbf{k_2}) e^{\frac{2\pi i \mathbf{k_2}\cdot\mathbf{r_2}}{L}}, \tag{C.29}$$

where $\mathbf{r_2} = (x,y)$ and $\mathbf{k_2} = (k_x, k_y)$.

Substituting equation C.26 in equation C.27 yields

$$\mathbf{F}_p(x,y) = \frac{1}{L^4}\int_{-L/2}^{L/2} dz \sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\sum_{k_z=-\infty}^{\infty} \tilde{\mathbf{F}}_p(\mathbf{k}) e^{\frac{2\pi i \mathbf{k}\cdot\mathbf{r}}{L}}. \tag{C.30}$$

Remembering that the integral

$$\frac{1}{L}\int_{-L/2}^{L/2} e^{\frac{2\pi i k_z \cdot z}{L}} dz = \begin{cases} 1 & \text{when} \quad k_z = 0, \\ 0 & \text{when} \quad k_z \neq 0, \end{cases} \tag{C.31}$$

equation C.30 becomes

$$\mathbf{F}_p(x,y) = \frac{1}{L^3}\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty} \tilde{\mathbf{F}}(k_x,k_y,k_z=0) e^{\frac{2\pi i \mathbf{k}\cdot\mathbf{r}}{L}}. \tag{C.32}$$

Comparing the inverse transform of $\mathbf{F}_p$, equation C.29, with equation C.32 shows that

$$\tilde{\mathbf{F}}_p(k_x,k_y) = \frac{1}{L}\tilde{\mathbf{F}}(k_x,k_y,k_z=0), \tag{C.33}$$

the Fourier series of the projected field $\mathbf{F}_p$ is proportional to $\tilde{\mathbf{F}}$ when the plane $k_z = 0$ is considered.

The definition of the power spectrum of $\mathbf{F}$ as the squared modulus of its Fourier transform, $\mathbf{P}(\mathbf{k}) = \tilde{\mathbf{F}}(\mathbf{k})\tilde{\mathbf{F}}^*(\mathbf{k})$, suggests a relation similar to equation C.33 between $\mathbf{P}(\mathbf{k})$ and $\mathbf{P}_p(\mathbf{k})$, the power spectrum of the projected field. For this relation to hold, however, it must be assumed both that the power spectrum defined on the plane $k_z = 0$ be statistically representative of the full power spectrum and it can be fully described as a function of the wave vector $k = |\mathbf{k}|$ with no angular dependence (isotropy).

Let

$$\langle \mathbf{F} \rangle = \frac{1}{L^3} \int_{-L/2}^{L/2} \int_{-L/2}^{L/2} \int_{-L/2}^{L/2} \mathbf{F}(x, y, z)\, dx\, dy\, dz \tag{C.34}$$

and

$$\langle \mathbf{F}^2 \rangle = \frac{1}{L^3} \int_{-L/2}^{L/2} \int_{-L/2}^{L/2} \int_{-L/2}^{L/2} \mathbf{F}^2(x, y, z)\, dx\, dy\, dz \tag{C.35}$$

be the mean value and the mean square value (spatial averages) of $\mathbf{F}$, respectively. In terms of the Fourier transform C.25, the mean values become

$$\langle \mathbf{F} \rangle = \frac{1}{L^3} \tilde{\mathbf{F}}(0, 0, 0). \tag{C.36}$$

The variance σ^2 of $\mathbf{F}$ is then given by

$$\sigma^2 = \langle \mathbf{F}^2 \rangle - \langle \mathbf{F} \rangle^2. \tag{C.37}$$

Invoking the Parseval's theorem (Rayleigh's identity) for discrete Fourier transforms [2]

$$\int_{-L/2}^{L/2} X(t)^2\, dt = \frac{1}{L} \sum_{k=-\infty}^{\infty} |\, \tilde{X}(k)\, |^2 = \frac{1}{L} \sum_{k=-\infty}^{\infty} \tilde{X}(k) \tilde{X}^*(k) \tag{C.38}$$

equation C.35 becomes

[2] Loosely speaking, the Parseval's theorem states that the power (inner product of a function with itself) computed on its original domain equals the power of its transform in Fourier space (Plancherel, 1910).

$$\langle \mathbf{F}^2 \rangle = \frac{1}{L^6} \sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \sum_{k_z=-\infty}^{\infty} \tilde{\mathbf{F}} \tilde{\mathbf{F}}^* = \frac{1}{L^6} \sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \sum_{k_z=-\infty}^{\infty} \mathbf{P}. \tag{C.39}$$

The definition of variance C.37 can thus be restated as

$$\sigma^2 = \frac{1}{L^6} \left(\left(\sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \sum_{k_z=-\infty}^{\infty} \mathbf{P} \right) - \tilde{\mathbf{F}}^2(0,0,0) \right), \tag{C.40}$$

where the relation C.36 was used. Similarly, the variance of the projected field $\mathbf{F}_2$ is simply

$$\sigma_p^2 = \frac{1}{L^4} \left(\left(\sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \mathbf{F}_p^* \tilde{\mathbf{F}}_p \right) - \tilde{\mathbf{F}}_p^2(0,0) \right), \tag{C.41}$$

and by the relation expressed by equation C.33

$$\begin{aligned} \sigma_p^2 =& \frac{1}{L^6} \left(\sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \tilde{\mathbf{F}}_{k_z=0} \tilde{\mathbf{F}}_{k_z=0}^* - \tilde{\mathbf{F}}^2(0,0,0) \right) \\ =& \frac{1}{L^6} \left(\sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \mathbf{P}_{k_z=0} - \tilde{\mathbf{F}}^2(0,0,0) \right) \\ , \end{aligned} \tag{C.42}$$

where the abbreviation $X_{k_z=0} = X(k_x, k_y, k_z = 0)$ on $\mathbf{F}$, its complex conjugate $\mathbf{F}^*$ and its spectral power $\mathbf{P}$ was used.

One can now construct the ratio between the variances of the observed field and the original field

$$R = \frac{\sigma_p^2}{\sigma^2} = \frac{\left(\left(\sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \mathbf{P}_{k_z=0} \right) - \tilde{\mathbf{F}}^2(0,0,0) \right)}{\left(\left(\sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \sum_{k_z=-\infty}^{\infty} \mathbf{P} \right) - \tilde{\mathbf{F}}^2(0,0,0) \right)}. \tag{C.43}$$

One can easily adapt the results found so far to those cases in which a physical field is represented as a discrete series of measurements at fixed grid points. Data collected from observed or simulated quantities are usually in this form.

To discretise the expressions above, a scale ratio, λ is introduced. This parameter is defined for each side of a data cube (parallelepiped) or rectangular image as the ratio of the side's size to the pixel size. Thus a cube (with equal sides) has a size of λ^3 pixels, while a square image is λ^2 pixels. The spatial frequencies at which the Fourier transform are evaluated become $k = -\lambda/2+1, -\lambda/2+2, ..., -2, -1, 0, 1, 2..., \lambda/2-1, \lambda/2$ along each axis.

One can now derive the spectral power $\mathbf{P}$ of the three-dimensional field $\mathbf{F}$ through its observed two-dimensional projection (up to a constant of proportionality). Consider the projected field $\mathbf{F}_p$, calculate its power spectrum $\mathbf{F}_p(k_x, k_y)$ and construct and from it construct the azimuthally averaged power spectrum $\mathbf{P}_p(k)(k)$, where $k = sqrtk_x^2 + k_y^2$ is the wave-vector. Under the assumption of isotropy, the following relation holds

$$\mathbf{P}_{k_z=0}(k) \propto \mathbf{P}_p(k) \tag{C.44}$$

This relation allows to re-write C.43 as

$$R = \frac{\left(\left(\sum_{k_x=-\lambda/2+1}^{\lambda/2} \sum_{k_y=-\lambda/2+1}^{\lambda/2} \mathbf{P}_p\right) - \mathbf{P}_p(0)\right)}{\left(\left(\sum_{k_x=-\lambda/2+1}^{\lambda/2} \sum_{k_y=-\lambda/2+1}^{\lambda/2} \sum_{k_z=-\lambda/2+1}^{\lambda/2} \mathbf{P}_p\right) - \mathbf{P}_p(0)\right)}, \tag{C.45}$$

or in a more compact notation

$$R = \frac{\sum_{k\neq 0}^{2\mathrm{D},\lambda} \mathbf{P}_p(k)}{\sum_{k\neq 0}^{3\mathrm{D},\lambda} \mathbf{P}_p(k)}, \tag{C.46}$$

where

$$\sum_{k\neq 0}^{2\mathrm{D},\lambda} \mathbf{P}_p(k) = \left(\sum_{k_x=-\lambda/2+1}^{\lambda/2} \sum_{k_y=-\lambda/2+1}^{\lambda/2} \mathbf{P}_p(k) \right) - \mathbf{P}_p(0) \tag{C.47}$$

and

$$\sum_{k\neq 0}^{3\mathrm{D},\lambda} \mathbf{P}_p(k) = \left(\sum_{k_x=-\lambda/2+1}^{\lambda/2} \sum_{k_y=-\lambda/2+1}^{\lambda/2} \sum_{k_y=-\lambda/2+1}^{\lambda/2} \mathbf{P}_p(k) \right) - \mathbf{P}_p(0) \tag{C.48}$$

As only the power spectrum of the projected field appear in C.43, once R is calculated, one can derive the variance of the full three-dimensional field $\mathbf{F}$ as $\sigma^2 = \sigma_p^2/R$. Also notice that since the scale ratio λ is a finite quantity, the observed variance (the projected field) and the estimated variance of the full field are lower limits to the actual variances that would be obtained in the limit $\lambda \to \infty$. citeBrunt2010 discuss this point in detail.

The general method to derive the variance of the full field presented above is applicable if and only if the projected field is the line-of-sight-averaged projection defined in equation C.27. When the observed field is the line-of-sight integral of the original field (column density derived from a density field, for instance), the method can still be applied provided that $\mathbf{F}_p$ is expressed in normalised units. This form of $\mathbf{F}_p$ is obtained by dividing $\mathbf{F}_p$ by its mean value. Normalised units for density fields are discussed in section 2.6 of Brunt et al. (2010) and an example is given below.

C.3 Density fields, an example

Let ρ be a three-dimensional density field, and N be its column density. Since both ρ and N are positive everywhere over their domain definition, one can express them in normalised units, obtained by dividing them by their mean values (ρ_0 and N_0). These units comply with $\mathbf{F}$ and its projection $\mathbf{F}_p$. Without this normalisation, the column density, defined as the line-of-sight averaged projection of ρ, is scaled by the size of the domain side L, as it is the integral of ρ, rather its the average. In observations where column densities are usually obtained through optically thin spectral lines or extinction maps, L is required to convert the column density to the projected mean density. However, this quantity is not always known, especially when accurate distances are not available.

With the variance of the normalised column density σ^2_{N/N_0} and the angular-averaged power spectrum $P_{N/N_0}(k)$, equation C.43 returns the variance of the normalised density field σ^2_{ρ/ρ_0},

$$\sigma^2_{\rho/\rho_0} = \sigma^2_{N/N_0} \frac{\left(\sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \sum_{k_z=-\infty}^{\infty} f(k)\right) - f(0)}{\left(\sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} f(k)\right) - f(0)} \tag{C.49}$$

Observe that the distance does not enter C.49 for the calculation of σ^2_{ρ/ρ_0}. However, for observations at a fixed angular resolution, GMCs at different distances refer to different physical scales introducing a dependence on distance in the calculation. In principle, the reasoning that leads to equation C.49 holds for any positive-valued field such as temperature (Brunt et al., 2010).

C.4 Solenoidal and Compressive modes

Assume that $\mathbf{F}$ is C^2 and consider its Helmholtz decomposition. Taking the Fourier transform of $\mathbf{F}$, it can be shown that in frequency space equivalent relations hold Stewart (2011)

$$\tilde{\mathbf{F}}(\mathbf{k}) = \tilde{\mathbf{F}}_{\perp}(\mathbf{k}) + \tilde{\mathbf{F}}_{\parallel}(\mathbf{k}), \tag{C.50}$$

$$\mathbf{k} \cdot \tilde{\mathbf{F}}_{\perp} = 0, \tag{C.51}$$

$$\mathbf{k} \times \tilde{\mathbf{F}}_{\parallel} = 0. \tag{C.52}$$

The constructions above justify the the use of the "$\parallel$" and "$\perp$" subscripts to refer to the curl-free component and the the divergence-free component of $\mathbf{F}$. At each point $\mathbf{k}$, by equation C.51, $\tilde{\mathbf{F}}_{\perp}$ is perpendicular (transversal) to $\mathbf{k}$. While equation C.52 indicates that $\tilde{\mathbf{F}}_{\parallel}$ is parallel to $\mathbf{k}$.

If one can choose a frame of reference in which $\langle \mathbf{F} \rangle = 0$, then the variance of $\mathbf{F}$ (equation C.37) becomes equivalent to the spatial average of $\langle \mathbf{F}^2 \rangle$,

$$\sigma^2 = \langle \mathbf{F}^2 \rangle. \tag{C.53}$$

Assuming that the domain of $\mathbf{F}$ is a cube of size L, by the linearity of the Fourier transform and by Parseval's theorem can write

$$\sigma^2_{\mathbf{F}_\perp} = \frac{1}{L^6} \sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \sum_{k_z=-\infty}^{\infty} \tilde{\mathbf{F}}_\perp \tilde{\mathbf{F}}^*_\perp, \tag{C.54}$$

and

$$\sigma^2_{\mathbf{F}_\parallel} = \frac{1}{L^6} \sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \sum_{k_z=-\infty}^{\infty} \tilde{\mathbf{F}}_\parallel \tilde{\mathbf{F}}^*_\parallel. \tag{C.55}$$

C.5 Projections

Suppose that only information about a component of the field $\mathbf{F}$ is known, and, as is often the case in observational data (spectral lines, for instance), this component is directed along the line of sight. If one takes the line of sight to match the z-axis, the observed component breaks into its longitudinal and transversal parts as

$$F_z = F_{z\perp} + F_{z\parallel}. \tag{C.56}$$

Transforming into Fourier space, $F_z\hat{\mathbf{z}}$ becomes $\tilde{F}_z\hat{\mathbf{k}}_z$, thus F_z translates into the component of the transformed field $\tilde{\mathbf{F}}$ as a along the $\mathbf{k}_z$-direction.

Now consider equations C.51 and C.52 and the conditions they impose on the components of transformed field, $\tilde{F}_\perp$ and $\tilde{F}_\parallel$. For the dot product $\mathbf{k}\cdot\tilde{\mathbf{F}}_\perp = k_x\tilde{F}_{x\perp} + k_y\tilde{F}_{y\perp} + k_z\tilde{F}_{z\perp}$ to vanish, $\tilde{F}_{z\perp}$ must equal 0 along the k_z-axis ($k_x = k_y = 0$). On the ($k_z = 0$)-plane, the condition $\mathbf{k}\times\tilde{\mathbf{F}}_\parallel = 0$ becomes

$$k_y\tilde{F}_{z\parallel}\hat{\mathbf{k}}_x - k_x\tilde{F}_{z\parallel}\hat{\mathbf{k}}_y = 0, \tag{C.57}$$

implying that

$$\tilde{F}_{z\parallel} = 0, \tag{C.58}$$

thus $F_z = F_{z\perp}$ everywhere on this plane.

Assuming that $\mathbf{F}_\perp$ and $\mathbf{F}_\parallel$ are isotropic fields, i.e., their power spectrum can be entirely described as a function of the wave vector (see section C.2), then

$$\mathbf{P}_\perp = \tilde{\mathbf{F}}_\perp \cdot \tilde{\mathbf{F}}_\perp^* = F_{\perp 0}^2 f_\perp(k) \tag{C.59}$$

and

$$\mathbf{P}_\parallel = \tilde{\mathbf{F}}_\parallel \cdot \tilde{\mathbf{F}}_\parallel^* = F_{\parallel 0}^2 f_\parallel(k) \tag{C.60}$$

with $f_\perp(k)$ and $f_\parallel(k)$ being function that describe the power distributions and $F_{\perp 0}^2$ and $F_{\parallel 0}^2$ scaling factors.

Observe that the power distributions of the components of these fields are not isotropic themselves, but their structure follows a predictable pattern:

$$\tilde{F}_{z\parallel}\tilde{F}_{z\parallel}^* = \tilde{\mathbf{F}}_\parallel \cdot \tilde{\mathbf{F}}_\parallel^* \frac{k_z^2}{k^2}, \tag{C.61}$$

$$\tilde{F}_{z\perp}\tilde{F}_{z\perp}^* = \tilde{\mathbf{F}}_\perp \cdot \tilde{\mathbf{F}}_\perp^* \frac{k_x^2 + k_y^2}{2k^2}, \tag{C.62}$$

$$\tilde{F}_{x\parallel}\tilde{F}_{x\parallel}^* = \tilde{\mathbf{F}}_\parallel \cdot \tilde{\mathbf{F}}_\parallel^* \frac{k_x^2}{k^2}, \tag{C.63}$$

$$\tilde{F}_{x\perp}\tilde{F}_{x\perp}^* = \tilde{\mathbf{F}}_\perp \cdot \tilde{\mathbf{F}}_\perp^* \frac{k_y^2 + k_z^2}{2k^2}, \tag{C.64}$$

$$\tilde{F}_{y\parallel}\tilde{F}_{y\parallel}^* = \tilde{\mathbf{F}}_\parallel \cdot \tilde{\mathbf{F}}_\parallel^* \frac{k_y^2}{k^2} \tag{C.65}$$

and

$$\tilde{F}_{y\perp}\tilde{F}_{y\perp}^* = \tilde{\mathbf{F}}_\perp \cdot \tilde{\mathbf{F}}_\perp^* \frac{k_x^2 + k_z^2}{2k^2}. \tag{C.66}$$

Isotropy is restored if $\tilde{\mathbf{F}}_\perp \cdot \tilde{\mathbf{F}}^*_\perp = 2\tilde{\mathbf{F}}_\parallel \cdot \tilde{\mathbf{F}}^*_\parallel$, $\forall\ k$.

In general, the power spectra of observed fields are not fully isotropic. In certain cases, one can assume statistical isotropy with values of the power oscillating around those of a fully isotropic power spectrum. If statistical isotropy alone is considered, by construction, $\tilde{F}_{z\parallel}\tilde{F}^*_{z\parallel} = 0$ must still hold everywhere on the $(k_z = 0)$-plane, thus

$$\tilde{F}_z\tilde{F}^*_z = \tilde{F}_{z\perp}\tilde{F}^*_{z\perp} = \frac{1}{2}\tilde{\mathbf{F}}_\perp \cdot \tilde{\mathbf{F}}^*_\perp, \tag{C.67}$$

on this plan.

When F_z is spatially averaged along the line-of-sight (z-direction),

$$F_{z,p}(x, y) = \frac{1}{L}\int_{-L/2}^{L/2} F_z(x, y, z)\, dz, \tag{C.68}$$

as equation C.33 shows, the Fourier transform of the projection becomes

$$\tilde{F}_{z,p}(k_x, k_y) = \frac{1}{L}\tilde{F}_z(k_x, k_y, k_z = 0). \tag{C.69}$$

Thus, the Fourier transform of $F_{z,p}$ is proportional to a the $(k_z = 0)$-cut of the transformed $\tilde{\mathbf{F}}$ of the original field $\mathbf{F}$. By equation C.58, it follows that only the transversal part of the full field $\mathbf{F}$ determines the projected z-component $F_{z,p}$.

Writing out the power spectrum of $F_{z,p}$,

$$P_{z,p}(k_x, k_y) = \tilde{F}_{z,p}\tilde{F}^*_{z,p}(k_x, k_y) = \frac{1}{L^2}\tilde{F}_{z\perp}\tilde{F}^*_{z\perp}(k_x, k_y, k_z = 0) = \frac{1}{L^2}\tilde{\mathbf{F}}_\perp \cdot \tilde{\mathbf{F}}^*_\perp(k_x, k_y, k_z = 0), \tag{C.70}$$

(where equations C.67 and C.69 were used), one sees that $P_{z,p}$ is obtained from the power spectrum of the transverse component of the full field alone (provided it satisfies equations C.61 - C.66). Again, using Parceval' theorem, one can introduce the variance of $F_{z,p}$ in terms of its power spectrum,

$$\sigma^2_{F_{z,p}} = \frac{1}{L^4}\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\tilde{F}_{z,p}\tilde{F}^*_{z,p} = \frac{1}{L^6}\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\frac{\tilde{\mathbf{F}}_\perp\cdot\tilde{\mathbf{F}}^*_\perp}{2}, \tag{C.71}$$

where the second equality is obtained via equation C.70.

By Parseval's theorem and equation C.62, the variance of $F_{z,\perp}$ in the three-dimensional domain of $\mathbf{F}$ can be expressed as

$$\sigma^2_{F_{z\perp}} = \frac{1}{L^6}\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\sum_{k_z=-\infty}^{\infty}\tilde{F}_{z,p}\tilde{F}^*_{z,p} = \frac{1}{L^6}\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\tilde{\mathbf{F}}_\perp\cdot\tilde{\mathbf{F}}^*_\perp\frac{(k_x^2+k_y^2)}{2k^2}. \tag{C.72}$$

Assuming that equation C.59 holds, there is a way to compute the variance of $F_{z\perp}$ over the three-dimensional domain of the field from the variance of the observed component $F_{z,p}$:

$$\frac{\sigma^2_{F_{z\perp}}}{\sigma^2_{F_{z,p}}} = \frac{\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\sum_{k_z=-\infty}^{\infty}f_\perp(k)\frac{k_x^2+k_x^2}{k^2}}{\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}f_\perp(k)}. \tag{C.73}$$

Notice that the scaling factor appearing in equation C.59 is not essential for the ratio above; however, it should be noted that this factor must be considered in the calculation of the absolute variance $\sigma^2_{F_{z\perp}}$.

For an isotropic field,

$$\sigma^2_{F_{z\perp}} = \frac{1}{3}\sigma^2_{F_\perp}$$

so that one can write

$$\sigma^2_{F_{z\perp}} = \frac{2}{3}\sigma^2_{F_{z,p}} \tag{C.74}$$

If either the total z-variance $\sigma^2_{F_z}$ or the ratio of projected-to-total z-variance,$\sigma^2_{F_{z,p}}/\sigma^2_{F_z}$ is known, one can compute the fractional power in transversal modes as

$$\begin{aligned}\frac{\sigma^2_{F_{z\perp}}}{\sigma^2_{F_z}} &= \frac{\sigma^2_{F_{z,p}}}{\sigma^2_{F_z}} \frac{\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\sum_{k_z=-\infty}^{\infty} f_\perp(k)\frac{k_x^2+k_x^2}{k^2}}{\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty} f_\perp(k)} \\ &= \frac{2}{3}\frac{\sigma^2_{F_{z,p}}}{\sigma^2_{F_z}} \frac{\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty}\sum_{k_z=-\infty}^{\infty} f_\perp(k)}{\sum_{k_x=-\infty}^{\infty}\sum_{k_y=-\infty}^{\infty} f_\perp(k)}.\end{aligned} \tag{C.75}$$

In turn, this ratio equals the fractional power in transversal modes of the original field **F**

$$\frac{\sigma^2_{F_\perp}}{\sigma^2_{F_z}} \approx \frac{\sigma^2_{F_{z\perp}}}{\sigma^2_{F_z}}. \tag{C.76}$$

Thus, calculating the fraction of power in transversal modes requires no information on the longitudinal power spectrum.

C.6 Momentum density and the solendoidal fraction

Consider spectral line observations of the ISM. The change dI of the spectral line intensity provided by an optically thin isothermal medium with uniform excitation of density ρ along an infinitesimal path dz at position z is

$$dI(v) = e\rho\Phi(v - v_z(z))dz, \tag{C.77}$$

where e is a constant. The normalised profile function Φ is generally expressed as a Gaussian

$$\Phi(v - v_z) = \frac{1}{\sqrt{2\pi\sigma^2_{t,i}}} \exp\left(-\frac{(v - v_z)^2}{2\sigma^2_{t,i}}\right), \tag{C.78}$$

where the dispersion caused by thermal and instrumental line broadening is encoded by $\sigma_{t,i}$. For molecular emission, $\sigma_{t,i}$ is usually negligible in comparison to the overall velocity dispersion. In this scenario, $\Phi(v - v_z)$ can be approximated by a Dirac delta function $\delta(v - v_z)$ and equation C.79 describes the distribution of intensity weighted line-of-sight velocities (Falgarone et al., 1994; Ostriker et al., 2001).

Define now the spectral line intensity observed along a line of sight across a distribution of medium of length L as the integrated intensity,

$$I(x,y,z) = e\int_{-L/2}^{L/2} \rho(x,y,z)\Phi(v - v_z(x,y,z))\,dz. \tag{C.79}$$

With the approximation $\Phi(v-v_z) = \delta(v-v_z)$, and calculate the first moment of velocity as

$$\begin{aligned} W_1 &= \int_{-\infty}^{\infty} I(x,y,z)v\,dz \\ &= \int_{-\infty}^{\infty} e\,dv \int_{-L/2}^{L/2} \rho(x,y,z)\delta(v - v_z)v\,dz \\ &= e\int_{-L/2}^{L/2} \rho(x,y,z)v_z(x,y,z)\,dz \\ &= e\int_{-L/2}^{L/2} p_z(x,y,z)\,dz \\ &= eLp_{z,p}, \end{aligned} \tag{C.80}$$

where $p_z = \rho v_z$ was used to denote the component of the "momemtum" $\mathbf{p} = \rho\mathbf{v}$ along the z-axis, while

$$p_{z,p}(x,y) = \frac{1}{L}\int_{-L/2}^{L/2} p_z(,x,z)\,dz \tag{C.81}$$

is its line-of-sight projection.

Thus from the definition of the first observable moment of velocity, it follows that the density momentum field satisfies (up to constants) the 'spatial projection' condition of C.27. Notice that a velocity field alone would not satisfy this condition unless it is restricted to uniform densities (see Brunt et al. (2010) and Brunt & Federrath (2014)). Considering this, one can now examine equation C.75 substituting F_z and $F_{z,p}$ with p_z and $p_{z,p}$ respectively. To evaluate the ratio between the power in the transversal modes of the line-of-sight (z-) momentum density (variance of $p_{z\perp}$) and the power in the full line-of-sight component (variance of p_z) through equation C.75, one needs to work out the ratio $\sigma^2_{p_{z,p}}/\sigma^2_{p_z}$ (relative fraction of z-momentum power projected on the observation

field) and the angle averaged transversal power spectrum $f_\perp(k)$. The latter can be derived directly by considering the power spectrum's angular average of W_1 (equation C.80). The constants e and L could be obtained from the size of the observation field; however, it is more convenient to normalise them out.

The zeroth velocity moment is

$$W_0(x,y) = \int_{-\infty}^{\infty} I(x,y,v)\,dv = eL\rho_p(x,y) = eN \tag{C.82}$$

with

$$\rho_p(x,y) = \frac{1}{L}\int_{-L/2}^{L/2} \rho(x,y,z)\,dz \tag{C.83}$$

being the column density along the line of sight.

Now considering the spatial averages of W_0 and W_1 estimated in the frame of reference of W_0 ($\langle W_1\rangle = 0 \iff \langle p_{z,p}\rangle = 0$), one sees that

$$\frac{\langle W_1^2\rangle}{\langle W_0^2\rangle} = \frac{\sigma_{p_{z,p}}^2}{\langle \rho_p^2\rangle}. \tag{C.84}$$

To determine the ratio $\sigma_{p_{z,p}}^2/\sigma_{p_z}^2$ needed for the solution of equation C.75, one notices that as $\sigma_{p_{z,p}}^2/\langle\rho_p^2\rangle$ is the projected counterpart of $\sigma_{p_z}^2/\langle\rho^2\rangle$. Thus, if an estimate of $\sigma_{p_z}^2/\langle\rho^2\rangle$ was available, one could construct the ratio

$$\frac{\sigma_{p_{z,p}}^2/\langle\rho_p^2\rangle}{\sigma_{p_z}^2\langle/\rho^2\rangle} = \frac{\sigma_{p_{z,p}}^2}{\sigma_{p_z}^2}\frac{\langle\rho^2\rangle}{\langle\rho_p^2\rangle} = \frac{\sigma_{p_{z,p}}^2}{\sigma_{p_z}^2}\frac{\langle(\rho/\rho_0)^2\rangle}{\langle(N/N_0)^2\rangle}. \tag{C.85}$$

where the column density N, the mean column density $N_0 = \rho_0/L$ and the mean volume density ρ_0 were introduced with

$$\rho_p = N/L = \rho_0(N/\rho_0 L) = \rho_0(N/N_0).$$

Consider the terms in the nominator and denominator of C.89. Using the spatial average of the (squared) zeroth moment, $\langle(\mathbf{N/N_0})^\mathbf{2}\rangle$ becomes

$$\left\langle \left(\frac{N}{N_0} \right)^2 \right\rangle = \frac{\langle N^2 \rangle}{\langle N_0 \rangle^2} = \frac{\langle W_0^2 \rangle}{\langle W_0 \rangle^2}. \tag{C.86}$$

Observing that for the variances of ρ/ρ_0 and N/N_0 are related to their spatial averages as

$$\sigma^2_{\rho/\rho_0} = \langle (\rho/\rho_0)^2 \rangle - 1,$$

$$\sigma^2_{N/N_0} = \langle (N/N_0)^2 \rangle - 1,$$

one recovers $\langle (\rho/\rho_0)^2 \rangle$ through equation C.49. In this case, the angular averaged power spectrum $f(k)$ refers to the column density. This quantity can be derived from the power spectrum of the integrated intensity $\tilde{W}_0 \tilde{W}_0^*$ (up to a negligible normalisation constant).

With these results, equation C.85 can be recast as

$$\frac{\sigma^2_{p_{z,p}}}{\sigma^2_{p_z}} = \left[\frac{\sigma^2_{p_{z,p}}}{\langle \rho_p^2 \rangle} \right] \left[\frac{\langle (N/N_0)^2 \rangle}{(\rho/\rho_0)^2} \right] \left[\frac{\sigma^2_{p.z}}{\langle \rho^2 \rangle} \right]^{-1} \tag{C.87}$$

However, at this stage there is an alternative form of $\sigma^2_{p_{z,p}}/\sigma^2_{p_z}$ but an estimate of $\sigma^2_{p_z}/\langle \rho^2 \rangle$ is still missing. Writing this quantity out in full, using the definitions of variance and spatial average, one has

$$\frac{\sigma^2_{p.z}}{\langle \rho^2 \rangle} = \frac{\frac{1}{L^3} \int_{-L/2}^{L/2} dx \int_{-L/2}^{L/2} dy \int_{-L/2}^{L/2} p_z^2 \, dz}{\frac{1}{L^3} \int_{-L/2}^{L/2} dx \int_{-L/2}^{L/2} dy \int_{-L/2}^{L/2} \rho^2 \, dz} \tag{C.88}$$

Recalling the definition of momentum density, one can interpret equation C.88 as the velocity dispersion in the z-direction weighted by ρ^2:

$$\frac{\sigma^2_{p.z}}{\langle \rho \rangle} = \frac{\frac{1}{L^3} \int_{-L/2}^{L/2} dx \int_{-L/2}^{L/2} dy \int_{-L/2}^{L/2} \rho^2 v_z^2 \, dz}{\frac{1}{L^3} \int_{-L/2}^{L/2} dx \int_{-L/2}^{L/2} dy \int_{-L/2}^{L/2} \rho^2 \, dz} = \frac{\langle \rho^2 v_z^2 \rangle}{\langle \rho^2 \rangle}. \tag{C.89}$$

From the datacube, one can access the z-velocity dispersion weighted with ρ. This is attained through the second velocity moment

$$
\begin{aligned}
W_2(x,y) &= \int_{-\infty}^{\infty} I(x,y,z)v^2\,dv \\
&= \int_{-\infty}^{\infty} e\,dv \int_{-L/2}^{L/2} \rho(x,y,z)\delta(v-v_z)v^2\,dz \\
&= e\int_{-L/2}^{L/2} \\
&\rho(x,y,z)v_z^2(x,y,z)\,dz,
\end{aligned}
\tag{C.90}
$$

and the ratio of the spatial averages of W_2 and W_0

$$
\frac{\langle W_2\rangle}{\langle W_0\rangle} = \frac{\frac{e}{L^3}\int_{-L/2}^{L/2}dx\int_{-L/2}^{L/2}dy\int_{-L/2}^{L/2}\rho^2 v_z^2\,dz}{\frac{e}{L^3}\int_{-L/2}^{L/2}dx\int_{-L/2}^{L/2}dy\int_{-L/2}^{L/2}\rho\,dz} = \frac{\langle\rho^2 v_z^2\rangle}{\langle\rho\rangle}. \tag{C.91}
$$

Equipped with equation C.19, the measurable quantity $\frac{\langle\rho v_z^2\rangle}{\langle\rho\rangle}$ can be linked to $\frac{\langle\rho^2 v_z^2\rangle}{\langle\rho^2\rangle}$ as

$$
\frac{\sigma_{p_z}^2}{\langle\rho^2\rangle} = \frac{\langle\rho^2 v_z^2\rangle}{\langle\rho^2\rangle} = g_{21}\frac{\langle\rho v_z^2\rangle}{\langle\rho\rangle}. \tag{C.92}
$$

where the correction factor g_{21} is of order unity and ϵ (equation C.19) is a small, positive constant (Brunt & Federrath (2014) show through numerical simulations that ϵ depends on the Mach number). Thus it has been shown that one can compute $\langle\rho v_z^2\rangle/\langle\rho\rangle$ using the ratio $\langle W_2\rangle/\langle W_0\rangle$.

Using this result in equation C.87 gives

$$
\frac{\sigma_{p_{z,p}}^2}{\sigma_{p_z}^2} = \left[\frac{\sigma_{p_{z,p}}^2}{\langle\rho_p^2]\rangle}\right]\left[\frac{\langle(N/N_0)^2\rangle}{(\rho/\rho_0)^2}\right]\left[g_{21}\frac{\langle\rho v_z^2\rangle}{\langle\rho\rangle}\right]^{-1}, \tag{C.93}
$$

so that the **solenoidal fraction**, the relative fraction of z-momentum power in transversal modes (to the power in the full projected component) is

$$
R = \frac{\sigma_{p_{z\perp}}^2}{\sigma_{p_z}^2} = \frac{\sigma_{p_{z,p}}^2}{\sigma_{p_z}^2}\frac{\sum_{k_x=-k_{\max}}^{k_{\max}}\sum_{k_y=-k_{\max}}^{k_{\max}}\sum_{k_z=-k_{\max}}^{k_{\max}} f_\perp(k)\frac{k_x^2+k_y^2}{k^2}}{\sum_{k_x=-k_{\max}}^{k_{\max}}\sum_{k_y=-k_{\max}}^{k_{\max}} f_\perp(k)}. \tag{C.94}
$$

Here k_{max} is the greatest wavenumber observed in the k_x-, k_y-, and k_z-directions and $f_\perp(k)$ the angular average of the projected momentum power spectrum. For isotropic power spectra, equation C.94 yields

$$\frac{\sigma^2_{p_\perp}}{\sigma^2_p} \approx \frac{\sigma^2_{p_{z\perp}}}{\sigma^2_{p_z}}. \tag{C.95}$$

C.7 Summary

One can study the momentum density constructed as $\mathbf{p} = \rho\mathbf{v}$ (see section C.6) from the volume density ρ and velocity $\mathbf{v}$ fields. From the line-of-sight projected transversal component of the momentum density, one can derive the relative ratio of power possessed by the solenoidal modes of the field (Helmholtz decomposition):

$$R = \frac{\sigma^2_{p_\perp}}{\sigma^2_p}. \tag{C.96}$$

R is referred to as the solenoidal fraction of the momentum density.

Assuming that the emission lines from ^{13}CO are optically thin and that emissivity only depends on the ^{13}CO molecular density, position-position-velocity data can be interpreted as a density-weighted velocity field. In this framework, the spectrum observed at a line of sight is the projection of the emission from the distribution of molecules along the line of sight, moving at different velocities. In a position-position-velocity datacube, velocity-weighted moments and their power spectra are available, directly measurable quantities. Via equations C.93, C.94, and C.95 the solenoidal fraction can be expressed with respect to these observables,

$$R = \left[\frac{\langle W_1^2\rangle}{\langle W_0^2\rangle}\right]\left[\frac{\langle W_0^2\rangle/\langle W_0\rangle^2}{1 + A(\langle W_0^2\rangle^2 - 2))}\right]\left[g_{21}\frac{\langle W_2\rangle}{\langle W_0\rangle}\right]^{-1}, \tag{C.97}$$

with

$$A = \frac{\left(\sum_{k_x=-k_{\max}}^{k_{\max}}\sum_{k_y=-k_{\max}}^{k_{\max}}\sum_{k_z=-k_{\max}}^{k_{\max}} f(k)\right) - f(0)}{\left(\sum_{k_x=-k_{\max}}^{k_{\max}}\sum_{k_y=-k_{\max}}^{k_{\max}} f(k)\right) - f(0)}, \tag{C.98}$$

$$B = \frac{\sum_{k_x=-k_{\max}}^{k_{\max}} \sum_{k_y=-k_{\max}}^{k_{\max}} \sum_{k_z=-k_{\max}}^{k_{\max}} f_{\perp}(k) \frac{k_x^2+k_y^2}{k^2}}{\sum_{k_x=-k_{\max}}^{k_{\max}} \sum_{k_y=-k_{\max}}^{k_{\max}} f_{\perp}(k)}, \tag{C.99}$$

and

$$f(k) = \frac{1}{2\pi k} \int_0^{2\pi} \tilde{W}_0(k,\Phi)\tilde{W}_0(k,\Phi)^* \, d\Phi, \tag{C.100}$$

$$f(k) = \frac{1}{2\pi k} \int_0^{2\pi} \tilde{W}_1(k,\Phi)\tilde{W}_1(k,\Phi)^* \, d\Phi, \tag{C.101}$$

being the angular average of the power spectra of the zeroth and first velocity moments of the line intensities, respectively.

The application of this method to segmented emission maps requires the clouds in the dataset to satisfy the isotropy and boundary conditions discussed above. The condition of statistical isotropy allows makes it possible to consider the projected two-dimensional averages to estimate the properties of the three-dimensional field. Individual filaments or clouds with strong anisotropy (due to strong magnetic field at low Mach numbers for instance Brunt & Federrath (2014)) must therefore be rejected. To avoid problematic boundary conditions, the momentum density is required to decay to zero smoothly at the cloud boundary. This condition assures a unique Helmholtz decomposition of the field into a solenoidal and compressible component and is necessary for the Fourier transform of the moments to be well-behaved (actual observed fields do not present periodic boundary conditions!). For segmentations in which the signal reaches the edges of the observation field, apodisation is necessary. Brunt et al. (2010) proved that their method is less accurate for fields that display steep power spectra. Such power distributions are sensitive to low spatial frequencies, which are often affected by uncertain statistics in the observation dataset (data affected by noise or the size of the telescope beam).

Appendix D

Random distance assignments

We construct three random distance assignment that consist of applying a distance to each SCIMES cloud by drawing the value

- from the set of unique distances that were assigned to SCIMES sources,
- from set of (equispaced) distances between the minimum and maximum value of the SCIMES distance assignments, distance assigned to SCIMES sources,
- from probability distribution (weights) generated from original distribution. of distances

The distances distributions derived from these assignment are compared to that of the original assignment (4) in Figure F.10.

Figure D.2 depicts the distribution of masses associated with the three random distance assignments.

The distributions of masses obtained through the random distance assignments are visually similar to the distribution generated with the original distances (see 4.2). These similarities suggest that, when a large sample of sources is considered, the distributions of quantities that depend on the cloud masses are not going to significantly impacted different distance assignment methods (see Chapter 4). Potentially, this observation may extend on all quantities that depend directly on distances.

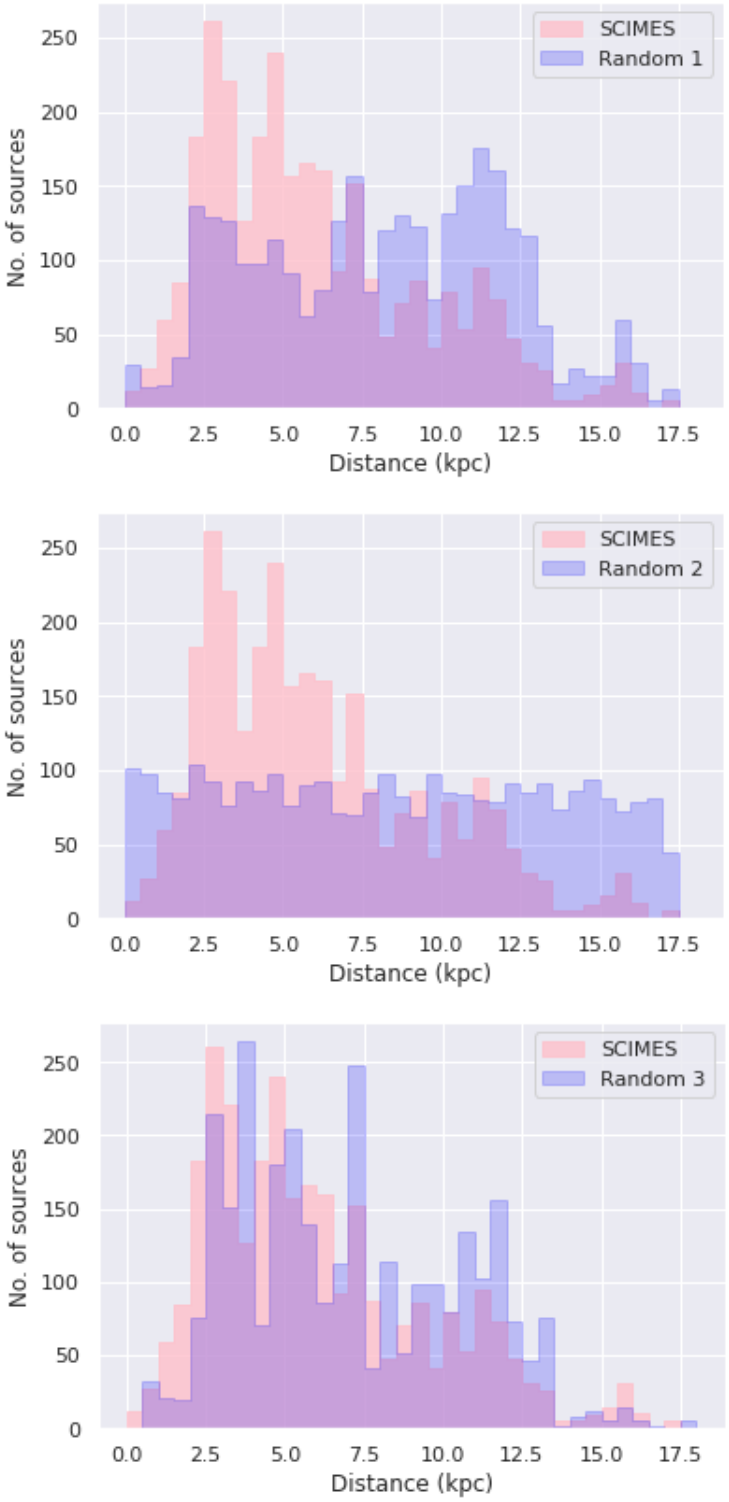


FIGURE D.1: Distribution of the three sets of random distances compared to the assigned distances to SCIMES clouds in CHIMPS (SCIMES). From top to bottom: the first set (Random 1) corresponds to distances drawn from the set of unique distances that were assigned to SCIMES sources The second set (Random 2) is drawn from set of (equispaced) distances between the minimum and maximum value of the SCIMES distance. Finally the set Random 3 is drawn from the distribution of distances generated from original SCIMES assignments.

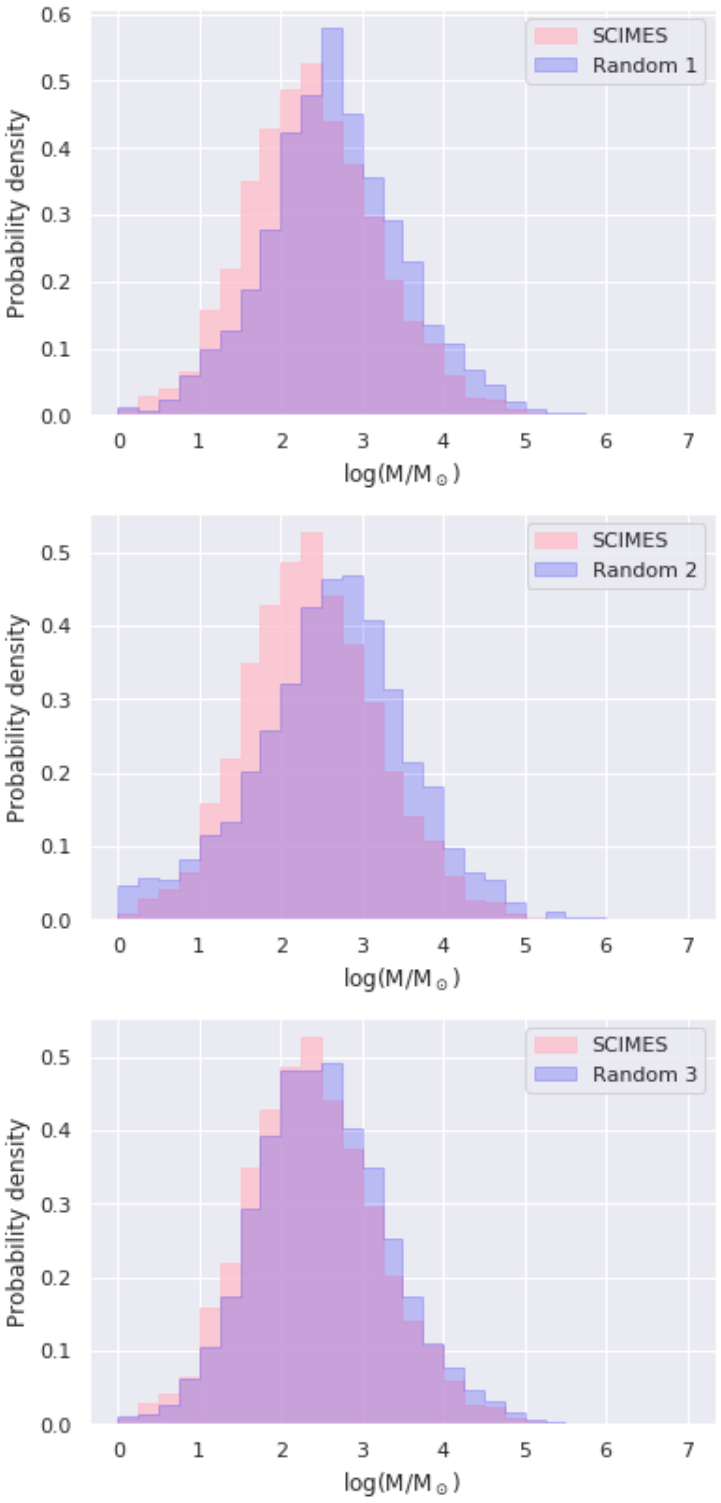


FIGURE D.2: Distribution of masses e estimated from the random distances sets (see figure F.10 compared to the masses corresponding to the original SCIMES distance assignments.

Appendix E

The FINDBACK filter

Findback is an application within the Starlink Kernel APplication PAckage (KAPPA) that estimates background noise in a datacube by removing small-scale structure[1]. We use Findback to subtract the background noise from the CHIMPS emission cubes. This step is crucial to both emission segmentation and the calculation of the solenoidal fraction (see Chapter 5).

The Findback filter consists of three subsequent searches in a cubic neighbourhood of each voxel. The size of the neighbourhood is specified as input and defines the scale of the smallest features not to be considered in the background estimate.

- **First pass:** The neighbourhood is searched for the minimum emission value. The filter then assigns it to the central voxel in the box.
- **Second pass:** The operation is repeated on the filtered data, this time replacing the central value with the maximum emission in the neighbourhood.
- **Third pass:** On the filtered data, the central value is substituted with the mean value in the neighbourhood.

The final mean-value surface provides an estimate of the 'lower envelope' of the data. This surface may present unnaturally sharp edges and it often follows the lower end of negative noise spikes. The latter problem leads to the underestimation of the true

[1] `http://starlink.eao.hawaii.edu/docs/sun255.htx/sun255ss4.htm`

background of the data. To remove sharp edges, the lower envelope is smoothed by re-applying the last step of the filter (mean-filter: substitution by neighbourhood's mean value).

Underestimation of the background and the fit of the lower envelope are addressed in several steps. First, the difference between the original data values and the background data is estimated in regions far from any bright source. Voxels with residuals that are larger than three times the RMS noise are given a 'bad' label. The good residuals are smoothed with a mean filter, and the bad ones are assigned values through the interpolation of the nearest good values. The residuals are extrapolated and extended to bright regions. They can thus be used as a background correction factor over the entire map. This correction surface is finally added onto the initial background estimate to obtain the final background that is then subtracted to the initial datacube.

Appendix F

FW distance assignments in SCIMES clouds

The ranges of the distances of the FW clouds contained in each SCIMES cloud in CHIMPS are plotted below.

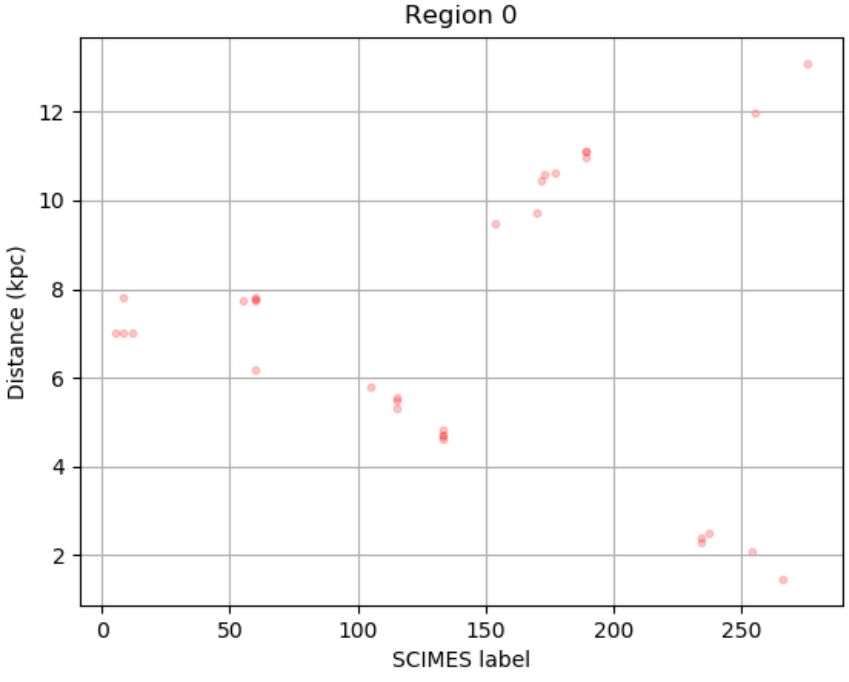


FIGURE F.1: FW distances assignments within SCIMES clouds in region 0.

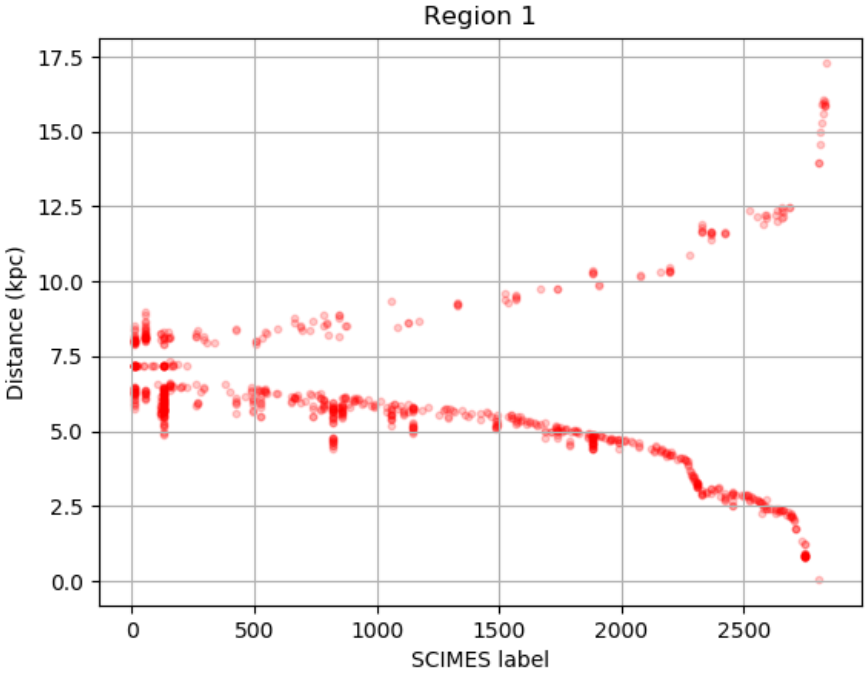


FIGURE F.2: FW distances assignments within SCIMES clouds in region 1.

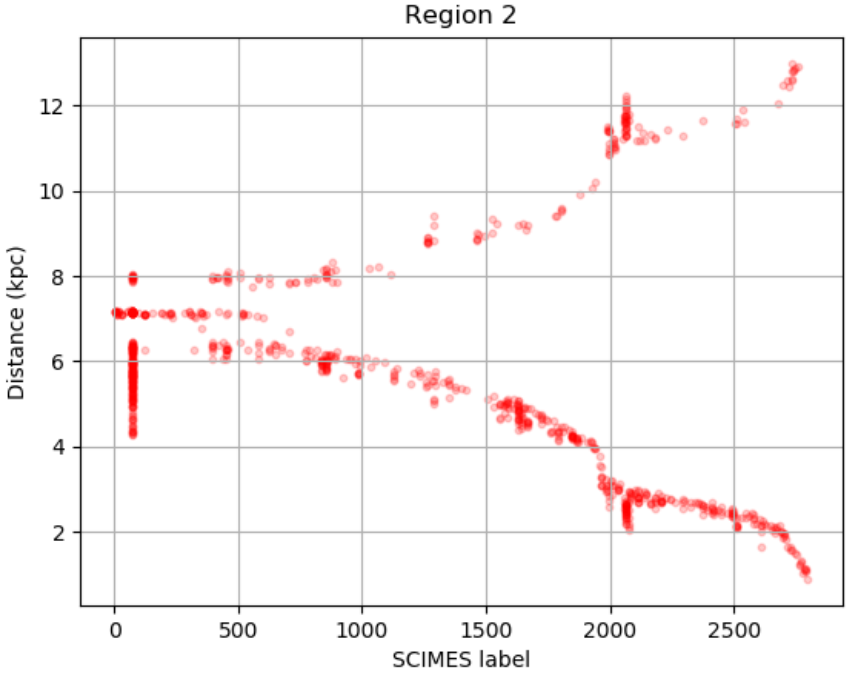


FIGURE F.3: FW distances assignments within SCIMES clouds in region 2.

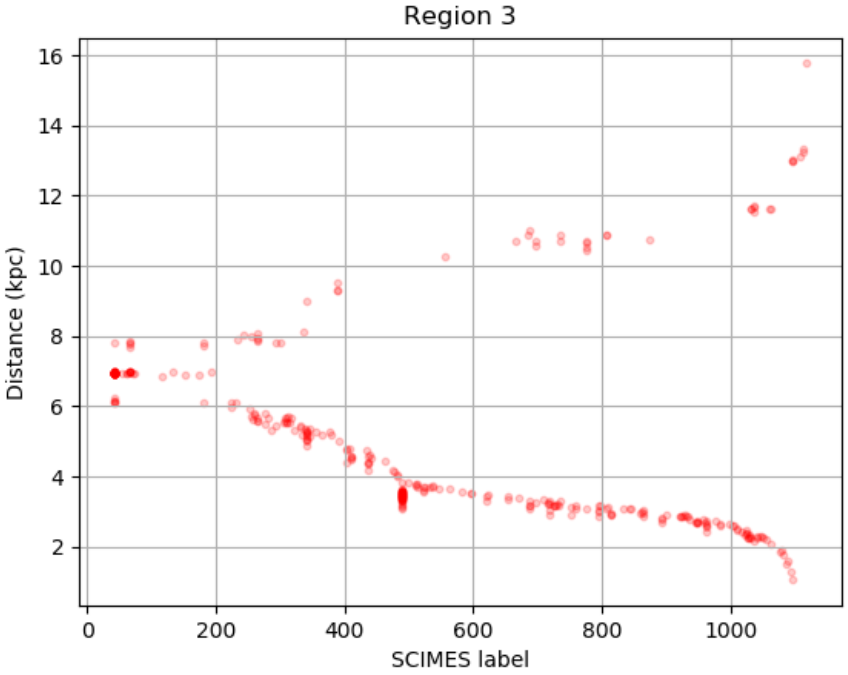


FIGURE F.4: FW distances assignments within SCIMES clouds in region 3.

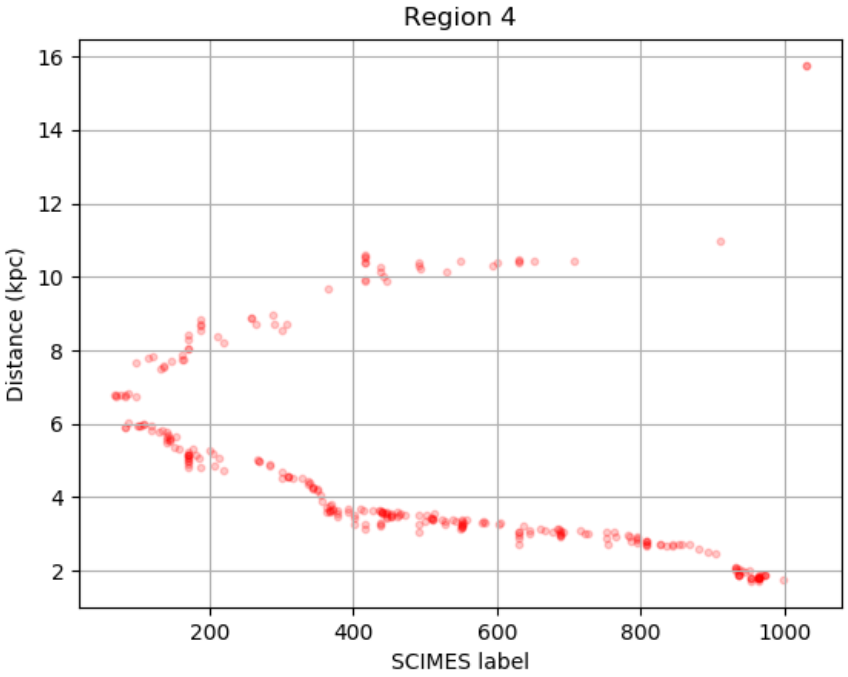


FIGURE F.5: FW distances assignments within SCIMES clouds in region 4.

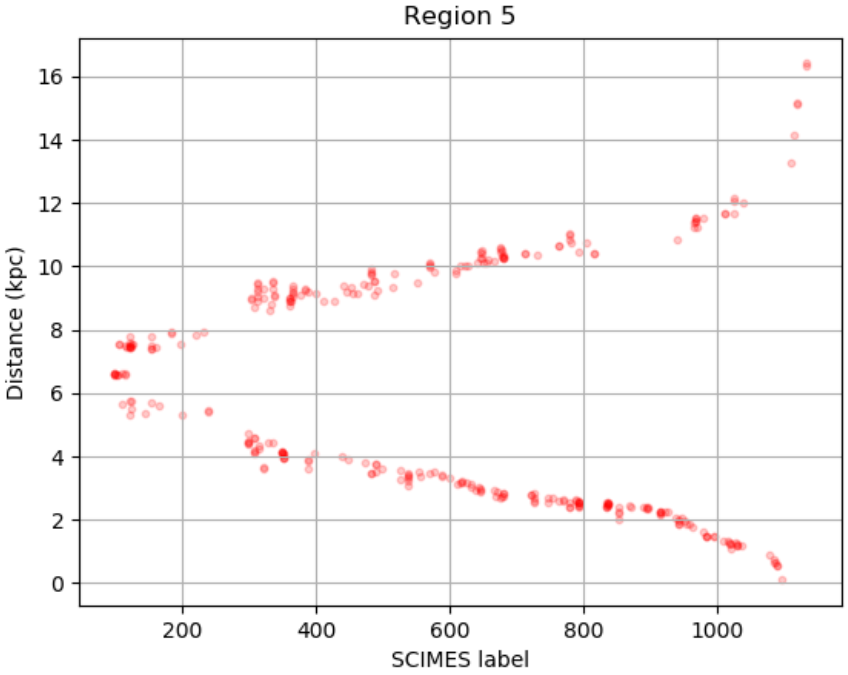


FIGURE F.6: FW distances assignments within SCIMES clouds in region 5.

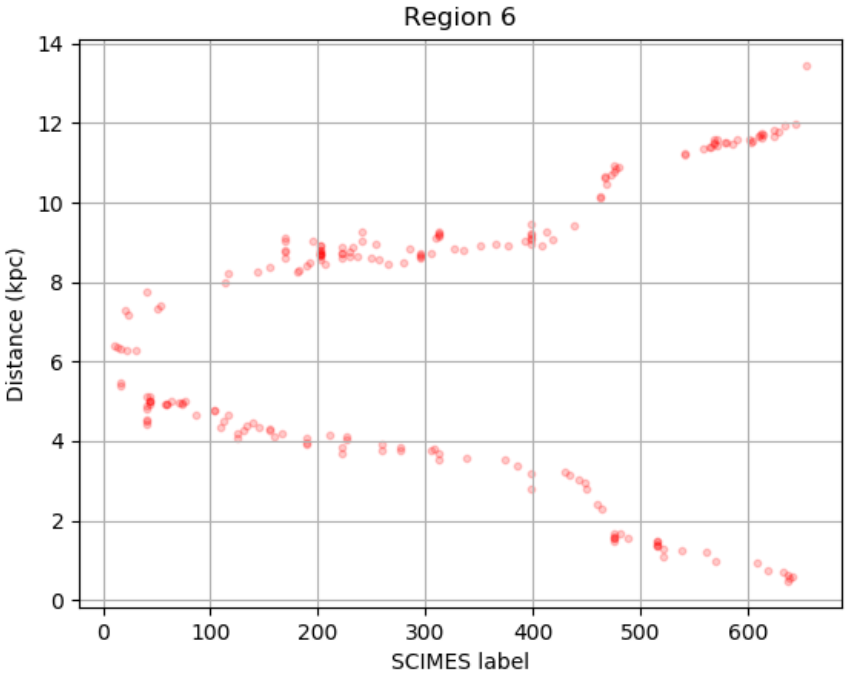


FIGURE F.7: FW distances assignments within SCIMES clouds in region 6.

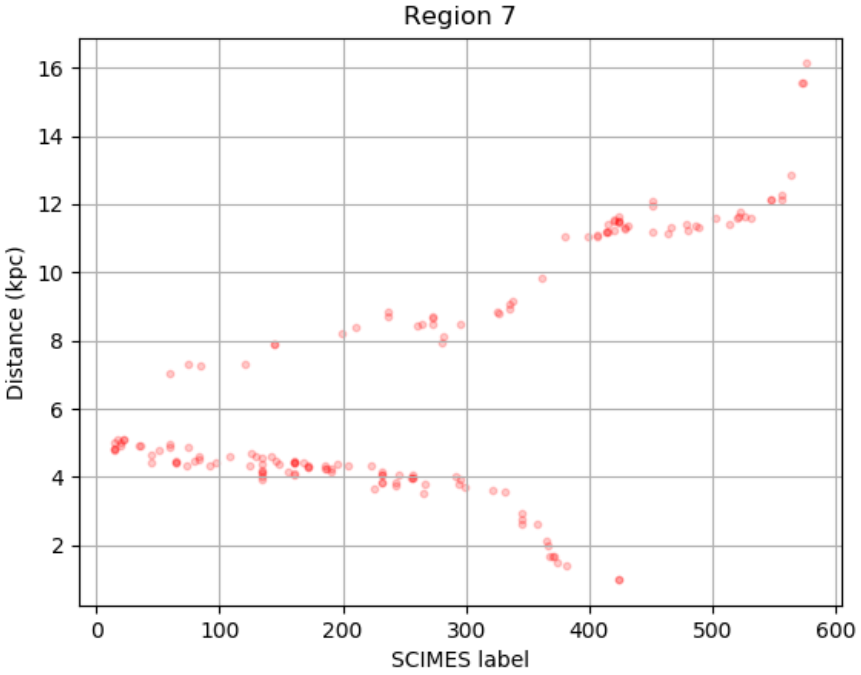


FIGURE F.8: FW distances assignments within SCIMES clouds in region 7.

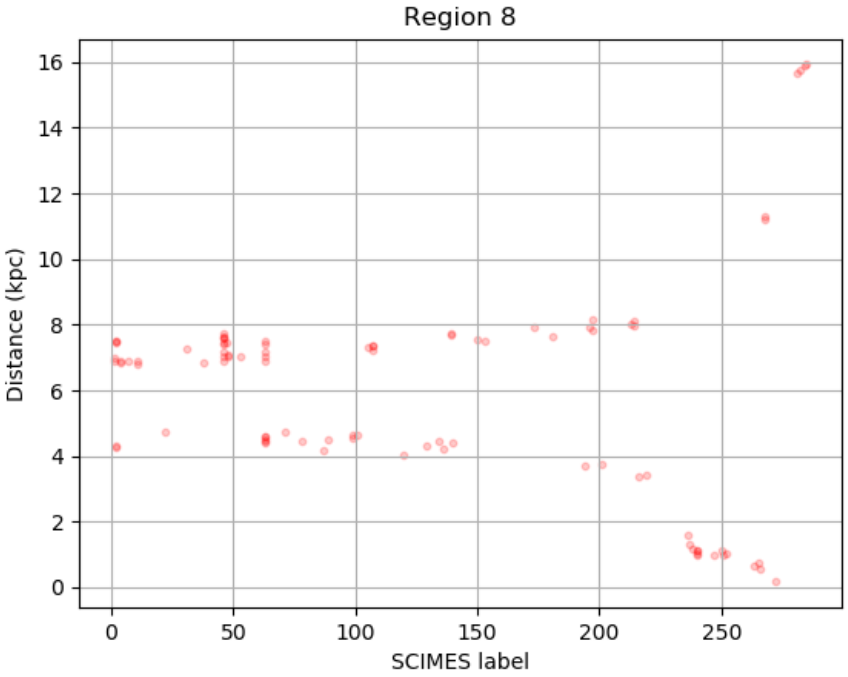


FIGURE F.9: FW distances assignments within SCIMES clouds in region 8.

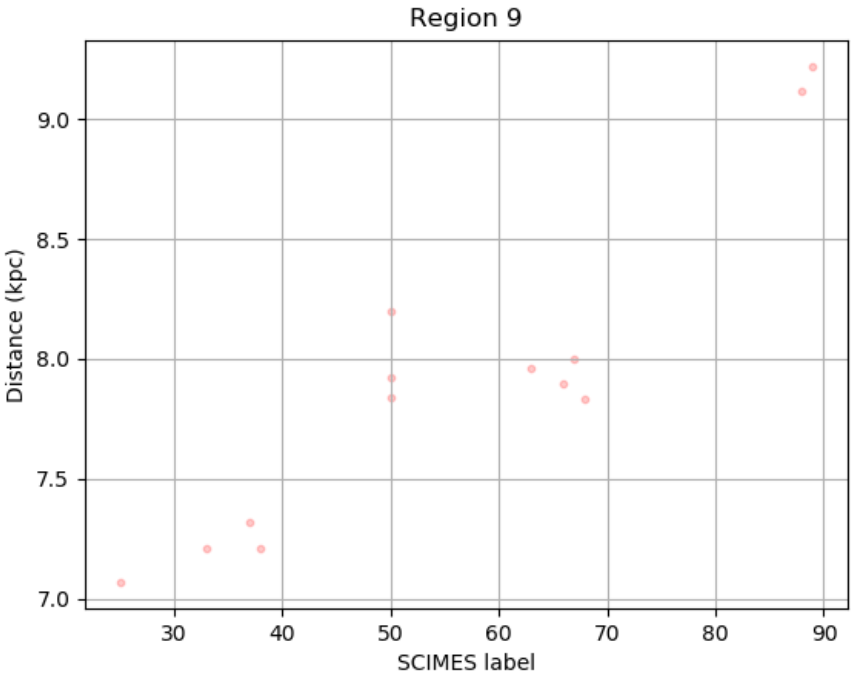


FIGURE F.10: FW distances assignments within SCIMES clouds in region 9.

Bibliography

Adamo A., et al., 2015, MNRAS, 452, 246

Aguerre J. E., et al., 2011, ApJS, 192, S82

Allen C. W., 1973, Astrophysical Quantities

Anderson L. D. B. T. M., 2009, ApJ, 690, 706

André P., et al., 2010, A&A, 518, L102

Arfken G. B., Weber H. J., 2005, Mathematical methods for physicists

Arthur D., Vassilvitskii S., , in Proc. 18th Annual ACM-SIAM Symp. on Discrete Algorithms, SODA '07

Arzoumanian D., et al., 2011, A&A, 529, L6

Aurenhammer F., 1991, ACM Computing Surveys, 23, 345

Ballesteros-Paredes J., Vázquez-Semadeni E., Scalo J., 1999, ApJ, 515, 286

Banerjee R., Pudritz R. E., 2006, ApJ, 641, 949

Barnes P. J., et al., 2015, ApJ, 812, 6

Bastian N., Covey K. R., Meyer M. R., 2010, ARA& A, 48, 339

Battersby C., et al., 2020, ApJS, 251, 35

Benedettini M., et al., 2020, A& A, 633, A147

Benjamin R. A., et al., 2005, ApJ, 630, L149

Bensch F., 2006, A&A, 448, 1043

Bensch F.and Stutzki J., Ossenkopf V., 2001, A & A, 366, 636

Bergin E. A., Tafalla M., 2007, ARA&A, 45, 339

Berry D. S., 2015, Astron. Comput., 10, 22

Bertin E., Arnouts S., 1996, A&AS, 117, 393

Beuther H., et al., 2012, ApJ, 747, 43

Beuther H., Klessen R., Dullemond C. P., et al., eds, 2014, Formation of molecular clouds and global conditions for star formation Springer

Bigiel F., Leroy A., Walter F., Brinks E., et al., 2008, AJ, 136, 2846

Blitz L., 1988, Millimetre and submillimetre astronomy. Kluwer, Dordrecht

Blitz L., 1991, The physics of star formation and early stellar evolution. Springer, Berlin

Blitz L., Fukui Y., Kawamura A., Leroy A., Mizuno N., Rosolowsky E., 2007, in Reipurth B., Jewitt D., Keil K., eds, , Protostars and Planets V. Univ. Arizona Press, Tucson

Blitz L., Stark A. A., 1986, ApJL, 300, L89

Bloemen J. B. G. M., 1985, A&A, 145, 391

Bloemen J. B. G. M., et al., 1984, A&A, 135, 12

Bohlin R. C., Savage B. D., Drake J. F., 1978, ApJ, 224, 132

Brand J., Blitz L., 1993, A&A, 275, 67

Brand J., Wouterloot J. G. A., 1995, A&A, 303, 851

Brand J., Wouterloot J. G. A., L R. A., de Geus E. J., 2001, A&A, 377, 641

Brunt C., Kerton C., Pomerleau1 C., 2003, ApJS, 144, S47

Brunt C. M., Federrath C., 2014, MNRAS, 442, 1451

Brunt C. M., Federrath C., Price d. J., 2010, MNRAS, 403, 1507

Buckle J. V., Hills R. E., Smith H., et al., 2009, MNRAS, 399, 1026

Burkhart B., Lazarian A., Ossenkopf V., Stutzki J., 2013, ApJ, 771, 123

Burton M., et al., 2013, Publ. Astron. Soc. Aust., 30, 44

Cavanagh B.and Jenness T., Economou F., Currie M. J., 2008, AN, 329, 295

Cheavance M., 2020, Space Sci. Rev., 216

Chung F. R. K., 1997, Spectral Graph Theory

Churchwell E., et al., 2009a, PASP, 121, 213

Churchwell E., et al., 2009b, PASP, 121, 213

Clemens D. P., 1985, ApJ, 295, 422

Clemens D. P., Barvainis R., 1988, ApJS, 68, 257

Clemens D. P., Yun J. L., H. H. M., 1991, ApJS, 75, 877

Colombo D., et al., 2014, ApJ, 784, 3

Colombo D., et al., 2019, MNRAS, 483, 4291

Colombo D., Rosolowsky E., Ginsburg A., Duarte-Cabral A., Hughes A., 2015a, MNRAS, 454, 2067

Colombo D., Rosolowsky E., Ginsburg A., Duarte-Cabral A., Hughes A., 2015b, MNRAS, 454, 2067

Combes F., 2012, A&A, 539, A67

Contreras Y., et al., 2013, A& A, 549, A45

Crowher P. A., et al., 2010, MNRAS, 408, 731

Csengeri T., et al., 2014, A& A, 565, A75

Currie M. J., 2013, in Friedel D. N., ed., Astronomical Society of the Pacific Conference Series, Vol. 475 Astronomical data analysis software and systems xxii. p. 341

Dame T. M., Hartmann D., Thaddeus 2001, ApJ, 547, 792

Davies B., et al., 2011, MNRAS, 416, 972

Dempsey J. T., Thomas H. S., Currie M. J., 2013, ApJs, 209, 8

di Francesco J., Evans N. J. I., Caselli P., Myers P. C., Shirley Y., Aikawa Y., Tafalla M., 2007, in Reipurth B., Jewitt D., Keil K., eds, , Protostars and Planets V. Univ. Arizona Press, Tucson

Dib S., 2014, MNRAS, 444, 1957

Dib S., Kim J., Shadmehri M., 2007, MNRAS, 381, 40

Dib S., Schmeja S., Hony S., 2017, MNRAS, 464, 1738

Dib S., Walcher C. J., Heyer M., Audit E., Loinard L., 2009, MNRAS, 398, 1201

Dickman R. L., Horvath M. A., Margulis M., 1990, ApJ, 365, 586

Dobbs C., Baba J., 2014, Publ. Astron. Soc. Aust., 31, e35

Dobbs C. L., Bonnell I. A., Pringle J. E., 2006, MNRAS, 371, 166

Dobbs C. L., Burkert A., Pringle J. E., 2011, MNRAS, 417, 1318

Draine B. T., 2011, Physics of the interstellar and intergalactic medium. Princeton University Press, Oxford

Duarte-Cabral A., Dobbs C. L., 2016, MNRAS, 458, 3667

Duarte-Cabral A., et al., 2021, MNRAS, 500, 3027

Dunham M. K., Rosolowsky E., Evans I. N. J., Cyganowski C., Urquhart J. S., 2011, APJ, 741, 110

Eden D., et al., 2015, MNRAS, 452, 289

Eden D. J., et al., 2013, MNRAS, 431, 15

Eden D. J., et al., 2020, MNRAS, 498, 5936

Eden D. J., et al., 2021, MNRAS, 500, 191

Eden D. J., Moore T. J. T., 2018, MNRAS

Eden D. J., Moore T. J. T., Plume R., et al., 2017, MNRAS, 469, 2163

Eden D. J., Moore T. J. T., Plume R., Morgan L. K., 2012, MNRAS, 422, 3178

Eisenstein D., Hut P., 1998, ApJ, 498, 137

Elia D., et al., 2007, ApJ, 655, 316

Elia D., et al., 2010, A&A, 518, L97

Elia D., et al., 2017, MNRAS, 471, 100

Elia D., et al., 2018, MNRAS, p. 509

Elia D., Molinari S., Fukui Y., et al., 2013, ApJ, 772, 45

Elia D., Strafella F., Schneider N., et al., 2014, ApJ, 788, 3

Elmegreen B. G., 1989, ApJ, 338, 178

Elmegreen B. G., 1993, ApJ, 411, 170

Elmegreen B. G., 2002, ApJ, p. 773

Elmegreen B. G., 2006, ApJ, 648, 572

Elmegreen B. G., 2007, ApJ, 668, 1064

Elmegreen B. G., 2008, ApJ, 672, 1006

Elmegreen B. G., Falgarone E., 1996, ApJ, 471, 816

Elmegreen B. G., Scalo J., 2004, ARA& A, 42, 211

Endres C. P., Schlemmer S., Schilke P., Stutzki J., Müller H. S. P., 2016, J. Mol. Spectrosc., 327, 95

Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96

Evans N. J. I., Heiderman A., Vutisalchavakul N., 2014, ApJ, 782, 114

Falgarone E., Lis D. C., Phillips T. G., Pouquet A., Porter D. H., Woodward 1994, ApJ, 436, 728

Falgarone E., Phillips T. G., Walker C. K., 1991, ApJ, 378, 186

Falgarone E., Puget J.-L., Pérault M., 1992, A& A, 257, 715

Feddersen J. R., Arce H. G., Kong S., Ossenkopf-Okada V., Carpenter J. M., 2019, ApJ, 875, 162

Federrath C., et al., 2016, ApJ, 832, 143

Federrath C., Klessen R., 2012, ApJ, 761, 156

Federrath C., Klessen R. S., Schmidt W., 2008a, ApJ, 688, 79

Federrath C., Klessen R. S., Schmidt W., 2008b, ApJ, 688, L79

Ferré-Mateu A., Vazdekis A., de la Rosa I. G., 2013, MNRAS, 431, 440

Fiedler M., 1973, Czech. Math. J., 23, 298

Field G. B., Blackman E. G., Keto E. R., 2011, MNRAS, 416, 710

Fischer I., Poland J., , 2004, New methods for spectral clustering

Fischer I., Poland J., 2005

Fitzpatrick E. L., 1999, PASP, 111, 63

Fontani F., et al., 2012, MNRAS, 423, 2342

Foyle K., Rix H.-W., Walter F., Leroy A. K., 2010, MNRAS, 725, 534

Fuller G. A., Peretto N., Pineda J. L., Molinari S., 2015, MNRAS, 451, 3089

Gao Y., Solomon P., 2004, ApJ, 606, 271

Gennaro M., et al., 2018, ApJ, 855, 20

Ghazzali N., Joncas G., Jean S., 1999, ApJ, 511, 242

Giannini T., et al., 2012, A&A, 539, A156

Gil de Paz A., et al., 2007, ApJS, 173, 185

Girichidis P., et al., 2018, MNRAS, 480, 3511

Goodman A. A., Barranco J. A., Wilner D. J., Heyer M. H., 1998, ApJ, 504, 223

Gray M. D., et al., 2016, MNRAS, 456, 374

Green J. A., McClure-Griffiths N. M., 2011, MNRAS, 417, 2500

Griffin M. J., Abergel A., Abreu et al., 2010, A&A, 518, L3

Grudič M., et al., 2018, MNRAS, 475, 3511

Haid S., et al., 2018, MNRAS, 478, 4799

Hartmann L., Ballesteros-Paredes J., Bergin E. A., 2001, ApJ, 562, 852

Hatchfield H. P., et al., 2020, ApJS, 251, 14

Haywood M., et al., 2009, A&A, 589, A66

Heitch F., et al., 2008, ApJ, 683, 786

Heitsch F., Hartmann L. W., Burkert A., 2008, ApJ, 683, 786

Helmholtz H., 1858, J. Reine Angew. Math., 55, 25

Henshaw J., et al., 2016, MNRAS, 457, 2675

Heyer M., et al., 2009, ApJ, 699, 1092

Heyer M. H., Brunt C. M., 2004, ApJL, 615, L45

Heyer M. H., Brunt C. M., 2012, MNRAS, 420, 1562

Heyer M. H., Terebey S., 1998, ApJ, 502, 265

Hily-Blant P., Teyssier D., Philipp S., Güsten R., 2005, ApJ, 440, 909

Hopkins A., 2013, MNRAS, 433, 170

Houlahan P.and Scalo J., 1990, ApJS, 72, 133

Houlahan P., Scalo J., 1992, ApJS, 393, 172

Hoyle F., 1953, ApJ, p. 513

Hughes A., et al., 2013, ApJ, 779, 46

Hunter J., Sandford M. T., Whitacker R. W., Klein R. I., 1986, ApJ, 305, 3

Inoue T., Inutsuka S. I., 2012, 2009, 704, 161

Izquierdo A. F., et al., 2021, MNRAS, 500, 5268

Jackson J. M., et al., 2006, ApJS, 163, S145

Jackson J. M., et al., 2013, Publ. Astron. Soc. Aust.,, 30, 57

Jain A. K., Murty M. N., Flynn P. J., 1999, Data clustering: A review

James P. A., Bretherton C. F., Knapen J. H., 2009, A&A, 501, 207

James P. A., Percival S. M., 2015a, MNRAS, 450, 3503

James P. A., Percival S. M., 2015b, MNRAS, 474, 3101

James P. A., Percival S. M., 2016, MNRAS, 457, 917

Jeffreson S. R., Kruijssen J. M. D., 2018, MNRAS, 476, 3688

Jenness T., Currie M. J., Tilanus R. P. J., Cavanagh B., Berry D. S., Leech J., Rizzi L., 2014, MNRAS, 453, 73

Jolliffe I. T., Cadima J., 2016, Philos. Trans. A Math. Phys. Eng. Sci., 374, 2065

Kainulainen J., Lada C., Rathborne J. M., Alves J. F., 2009, A&A, 497, 339

Kauffmann J., Pillai T., Goldsmith P. F., 2013a, ApJ, 779, 185

Kauffmann J., Pillai T., Goldsmith P. F., 2013b, ApJ, 779, 185

Kennicutt R., 1998, ApJ, 498, 541

Kennicutt R. C., et al., 2003, PASP, 115, 928

Khoperskov S., Haywood M., Di Matteo P., Lehnert M. D., Combes F., 2018, A&A, 609, A60

Kim J. G., Kim W. T., Ostriker E. C., 2018, ApJ, 859, 68

Kim W.-J., et al., 2017, A& A, 602, A37

Kirk H., et al., 2016, ApJ, 817, 167

Klein R. I., Woods T., 1998, ApJ, 497, 777

Klessen R., Heitsch F., Mac Low M. M., 2000, ApJ, 480, 887

Klessen R. S., Glover S. C. O., 2015, MNRAS, 451, 196

Kobayashi N., Yasui C., Tokunaga A. T., Saito M., 2009, ApJ, 683, 178

Koch E., et al., 2019, ApJ, 158, 1

Kolpak M. A., et al., 2003, ApJ, 582, 756

Könyver V., et al., 2015, A&A, 584, A91

Kornreich P., Scalo J., 2000, ApJ, 531, 366

Kramer C., Stutzki J., Rohrig R., Corneliussen U., 1998, A&A, 329, 249

Krause M., 2020, Space Sci. Rev., 216

Kroupa P., Weidner C., Pflamm-Altenburg J., Thies I., Dabringhausen J., Marks M., Maschberger T., 2013. Springer Netherlands, Dordrecht, p. 115

Kruijssen J., et al., 2014, MNRAS, 440, 3370

Kruijssen J., Longmore S., 2014, MNRAS, 439, 3239

Kruijssen J. M. D., 2012, MNRAS, 426, 3008

Kruijssen J. M. D., et al., 2019a, Nature, 569, 519

Kruijssen J. M. D., et al., 2019b, MNRAS, 484, 5734

Kruijssen J. M. D., Longmore S. N., 2013, ApJ, 435, 2598

Kruijssen J M. D., et al., 2019, Nature, 569, 519

Krumholz M. R., 2019, Phys. Rep., 539, 49

Krumholz M. R., Cunningham A. J., Klein R. I., McKee C. F., 2010, ApJ, 713, 1120

Krumholz M. R., Dekel A., McKee C. F., 2012, ApJ, 745, 69

Krumholz M. R., McKee C. F., 2005, ApJ, 630, 250

Krumholz M. R., McKee C. F., Bland-Hawthorn J., 2019, ARA& A, 57, 227

Krumholz M. R., McKee C. F., Tumlinson J., 2009, ApJ, 669, 850

Ktumholz M. R., 2014, Phys. Rep., 539, 49

Kutner M. L., Ulich B. L., 1981, ApJ, 250, 341

L. C. N., et al., 2011, ApJ, 741, 21

Lada C., et al., 2012, ApJ, 745, 190

Lada C. J., Lada E. A., 2003, ARA&A, 41, 57

Lada C. J., Muench A. A., Rathborne J. M., Alves J. F., Lombardi M., 2008, ApJ, 672, 410

Lada E. A., 1992, ApJL, 393, L25

Lada J. C., 2010, ApJ, 724, 687

Lada J. C., Dame T. M., 2020, ApJ, 898, 3

Langer W. D., Penzias A. A., 1990, A&A, 357, 477

Larson R. B., 1981, MNRAS, 194, 809

Lazarian A., Pogosyan D., 2000, ApJ, 537, 720

Lazarian A., Pogosyan D., 2004, ApJ, 616, 965

Lee Y., et al., 2016, J. Korean Astron. Soc., p. 255

Lépine J. R. D., et al., 2011, MNRAS, 417, 698

Leroy A. K., et al., 2009, AJ, 137, 4670

Leroy A. K., et al., 2013, ApJ, 146, 19

Leroy m. A. K., Walter F., Brinks E., Bigiel F., de Blok W. J. G., Madore B., Thornley M. D., 2010, AJ, 136, 2782

Li D., Goldsmith P. F., 2003a, ApJ, 585, 823

Li D., Goldsmith P. F., 2003b, ApJ, 585, 823

Li G.-X., Buckert A., Megeath T., Wyrowski F., Shi X., 2016

Li G.-X., Burkert A., 2017, MNRAS, 464, 4096

Lleti R., Ortiza M. C., Sarabia L. A., Sánchez S., 2004, Anal. Chim. Acta, 515, 87

Longmore S., et al., 2013, MNRAS, 429, 987

Lopez L. A., et al., 2014, ApJ, 795, 121

Mac Low M.-M., Klessen R. S., 2004, Rev. Mod. Phys., 76, 125

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

McKee C. F., Williams J. P., 1997, ApJ, 476, 144

MacLaren I., Richardson K. M., Wolfendale A. W., 1988, ApJ, 333, 821

McLeod A. F., et al., 2020, ApJ, 891, 25

MacQueen J., , in Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability Vol. 1, Statistics

Maíz Apellániz J., Úbeda L., 2005, ApJ, 629, 873

Meidt S. E., et al., 2013, ApJ, 779, 45

Meidt S. E., et al., 2018, ApJ, 854, 100

Meidt S. E., et al., 2019, ApJ, 892, 73

Meidt S. E., et al., 2020, ApJ, 892, 73

Men'shchikov et al., 2012, A&A, 542, A81

Men'shchikov A., André P., Didelon P., 2010, A&A, 518, L103

Meurer G. R., et al., 2009, ApJ, 695, 765

Milam S. N., Savage C., Brewster M. A., 2005, ApJ, p. 1126

Miville-Deschênes M.-A., Murray N., Lee E. J., 2017, ApJ, p. 57

Moisés A. P., et al., 2011, MNRAS, 411, 705

Molinari S., et al., 2010a, PASP, 122, 314

Molinari S., et al., 2010b, A&A, 518, L100

Molinari S., et al., 2011, A& A, 530, A133

Molinari S., et al., 2016, A& A, 591, a149

Molinari S., Pezzuto S., Cesaroni R., Brand J., Faustini F., Testi L., 2008, A&A, 481, 345

Moore T. J., et al., 2007, MNRAS, 379, 663

Moore T. J. e. a., 2012, MNRAS, 426, 7017

Moore T. J. T., et al., 2015, MNRAS, 453, 4264

Motte F., Schilke P., Lis D. C., 2003, ApJ, 582, 277

Mottram J., et al., 2011, ApJL, 730, L33

Muders D., Hafok H., Wyrowski F., et al., 2006, A&A, 454, L25

Myers P. C., Ladd E. F., 1993, ApJL, 413, L47

Nakamura F., Li Z. Y., 2007, ApJ, 662, 395

Narayanan D., Davé R., 2012, MNRAS, 423, 3601

Ng A. Y., Jordan M. I., Weiss Y., 2001, Advances in the neural information processing systems

Obreschkow D., Rawlings S., 2009, MNRAS, 394, 1857

Offner S. S. R., Arce H. G., 2015, ApJ, 811, 146

Onus A., Krumholz M. R., Federrath C., 2018, MNRAS, 479, 1702

Orkisz J. H., et al., 2017, A&A, 599, A99

Ostriker E. C., Stone J. M., Gammie C. F., 2001, ApJ, 546, 980

Padoan P., Federrath C., Chabrier G., Evans N. J. I., Johnstone D., Jørgensen J. K., McKee C. F., Nordlund A., Protostars and planets vi., booktitle = , year = 2013, publisher = , editor = Beuther, H. and Klessen, R. S. and Dullemond C. P. and Henning, T., pages = 77,

Padoan P., Goodman A. A., Juvela M., 2003, ApJ, 588, 881

Padoan P., Nordlund A., 2011, ApJ, 730, 40

Padoan P., Nordlund A., Jones B. J. T., 1997, MNRAS, 288, 145

Pal N. R., Pal S. K., 1993, Pattern Recognit., 26, 1277

Paradis D., Dobashi K., Shimoikura T., Kawamura A., Onishi T., Fukui Y., Bernard J.-P., 2012, A&A, 543, A103

Peretto N., Fuller G. A., 2009, A&A, 505, 405

Pety J., et al., 2017, A&A, 599, A98

Piazzo L., Ikhenaode D., Natoli P., et al., 2012, TIP, 21, 3687

Pineda J. E., Rosolowsky E. W., Goodman A. A., 2009, ApJ, 699, L134

Pingel N. M., Lee M.-Y., Burkhart B., Stanimirović S., 2018, ApJ, 856, 136

Plancherel M., 1910, Rendiconti del Circolo Mat. di Palermo, 30, 298

Poglitsch A., Waelkens C., Geis N., et al., 2010, A&A, 518, L2

Polychroni D., et al., 2013, APJL, 777, L33

Pon A., et al., 2012, ApJ, 756, 145

Pon A., et al., 2014, ApJ, 756, 145

Pon A., et al., 2016, A&A, 587, A96

Pound M. W., Goodman A. A., 1997, ApJ, 482, 334

Ragan S., et al., 2016, MNRAS, 462, 3123

Ragan S. E., et al., 2018, MNRAS, 479, 2361

Ragan S. E., Henning T., Beuther H., 2013, A&A, 559, A79

Rahner D., et al., 2017, MNRAS, 483, 4453

Rahner D., et al., 2019, MNRAS, 483, 2547

Rathborne J. M., et al., 2009, ApJS, 182, 131

Rathborne J. M., et al., 2014, ApJL, 795, L25

Rathborne J. M., M. J. J., R. S., 2006, ApJ, 641, 389

Reed B. C., 2000, AJ, 120, 314

Reid M. A., Wadsley J., Petitclerc N., Sills A., 2010, ApJ, 719, 561

Reid M. J., Dame T. M., Menten K. M., Brunthaler A., 2016, ApJ, 823, 77

Reid M. J., et al., 2014, ApJ, 783, 130

Reina-Campos M., Kruijssen J. M. D., 2017, MNRAS, 469, 1282

Reynaud D., Downes D., 1998, A&A, 337, 671

Rice T. S., et al., 2020, ApJ, 822, 52

Rigby A. J., Moore T. J. T., Eden D. J., Uruqhart J. S., et al., 2019, A&A

Rigby A. J., Moore T. J. T., Plume R., et al., 2016, MNRAS, 456, 2885

Roerdink J., Meijster A., 2001, Fund. Inform., 41, 187

Roman-Duval J., et al., 2009, ApJ, 699, 1153

Roman-Duval J., Jackson J. M., Heyer M., Rathborne J., Simon R., 2010, ApJ, 723, 492

Rosolowsky E., Leroy A., 2006, PASP, 118, 590

Rosolowsky E. W., Goodman A. A., Wilner D. J., Williams J. P., 1999, ApJ, 524, 887

Rosolowsky E. W., Pineda J. E., Kauffmann J., Goodman A. A., 2008, ApJ, 679, 1338

Rosseeuw P., 1987, J. Comput. Appl. Math., 20, 53

Roueff A., et al., 2020, A&A, p. A26

Rudolph A. L., Simpson J. P., Haas M. R., Erickson E. F., Fich M., 1997, ApJ, 489, 94

Salpeter E. E., 1955, ApJ, 121, 161

Sánchez N., Alfaro E. J., Pérez E., 2005, ApJ, p. 849

Sanders D. B., Scoville N. Z., Solomon P. M., 1985, ApJ, 289, 373

Saraceno P.and Andre P., Ceccarelli C., Griffin M., Molinari S., 1996, A&A, 309, 827

Scalo J. B. G., Elmegreen 2004, ARA& A, 42, 275

Schlegel D. J., Finkbeiner D. P., 1998, ApJ, 500, 525

Schöier F. L., et al., 2005, A& A, 432, 369

Schruba A., et al., 2011, AJ, 142, 37

Schruba A., Kruijssen J. M. D., Leroy A. K., 2019, ApJ, 883, 2

Schuller F., 2012, in Holland W. S., ed., Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy VI Vol. 8452, BoA: a versatile software for bolometer data reduction. SPIE

Schuller F., et al., 2009, A& A, 504, 415

Schuller F., et al., 2017, A&A, 601, A124

Scoville N. Z., Yun M. S., Sanders D. B., Clemens D. P., Waller W. H., 1987, ApJS, 63, 821

Shetty R., et al., 2012, MNRAS, 425, 720

Shi J., Malik J., 2000, IEEE Trans. Pattern Anal. Mach. Intell., 22, 888

Smith M. D., 2014, MNRAS, 438, 1051

Smith R., Clark P., Bonnell I. A., 2008, MNRAS, 391, 1091

Smith R. J., et al., 2020, MNRAS, 492, 1594

Sodroski T. J., et al., 1995, ApJ, 452, 262

Sofue Y., Nakanishi H., 2016, PASJ, 68, 63

Solomon P. M., Rivolo A. R., Barrett J., A. Y., 1987, ApJ, 319, 730

Sormani M. C., et al., 2019, MNRAS, 488, 4663

Spinoso D., et al., 2017, MNRAS, 465, 3729

Sreenivasan K., 1991, Ann. Rev. Fluid Mech., 23, 539

Stewart A. M., 2011, Sri Lankan J. Phys., 12, 33

Still S., Bialek W., 2004, Neural Comput., 16, 2483

Stutzki J., Bensch F., Heithausen A., Ossenkopf V., Zielinsky M., 1998, A&A, 336, 697

Stutzki J., Güsten R., 1990, ApJ, 356, 513

Sun J., et al., 2018, ApJ, 860, 172

Sun J., et al., 2020, ApJ, 892, 148

T. C., et al., 2016, A&A, 586, A149

Tafalla M., Myers P. C., Caselli P., Walmsley C. M., 2004, A&A, 416, 191

Teyssier R., 2002, A& A, 385, 337

Tibshirani R., Walther G., Hastie T., 2001, J. R. Statist. Soc. B, 63, 411

Toalà J. A., Vàzquez-Semadeni E., Gòmez G. C., 2012, ApJ, 744, 190

Traficante A., Calzoletti L., Veneziani M., et al., 2011, MNRAS, 416, 2932

Tubbs A. D., 1982, ApJ, 255, 458

Umemoto T., et al., 2017, PASJ, 69, 1

Urquhart J. S., et al., 2007, A&A, 474, 891

Urquhart J. S., et al., 2008, A& A, 487, 253

Urquhart J. S., et al., 2011, MNRAS, 418, 1689

Urquhart J. S., et al., 2012, MNRAS, 420, 1656

Urquhart J. S., et al., 2013, MNRAS, 431, 1752

Urquhart J. S., et al., 2014a, A& A, 568, A41

Urquhart J. S., et al., 2014b, MNRAS, 437, 179

Urquhart J. S., et al., 2018, MNRAS, 473, 1059

Vàzquez-Semadeni E., 1994, ApJ, 423, 681

Vàzquez-Semadeni E., A. G.-S., Colìn P., 2017, MNRAS, 467, 1318

Vàzquez-Semadeni E., et al., 2011, MNRAS, 414, 2511

Vishniac E., 1994, ApJ, 428, 186

von Luxburg U., 2007, Stat. Comp., 17, 395

Walker D. L., 2021, MNRAS, 503, 77

Walter F., et al., 2008, AJ, 136, 2563

Ward-Thompson D., André P., Crutcher R., Johnstone D., Onishi T., Wilson C., 2007, in Reipurth B., Jewitt D., Keil K., eds, , Protostars and Planets V. Univ. Arizona Press, Tucson

Weidner C., Kroupa P., 2006, MNRAS, 365, 1333

West A., 2002, Graph theory

Wienen M., et al., 2012, A&A, 544, A146

Wienen M., et al., 2015, A&A, 579, A91

Williams J. P., de Geus E. J., Blitz L., 1994, ApJ, 428, 693

Wilson B. A., Dame T. M., Masheder M. R. W., Thaddeus P., 2005, A&A, 430, 523

Wilson T. L., Rohlfs K., Hüttemeister S., 2013, Tools for Radio Astronomy. Springer, Berlin

Wilson T. L., Rood R., 1994, ARA&A, 32, 191

Wolfire M., et al., 2003, ApJ, 587, 278

Wong T., et al., 2011, ApJS, 197, S16

Wouterloot J. G. A., Brand J., Burton W. B., Kwee K. K., 1990, A&A, 230, 21

Zavagno A., Anderson L. D., Russeil D., et al., 2010, A&A, 518, L101

Zelnik-Manor L., Perona P., 2004, in dvances in Neural Information Processing Systems 17

Zetterlund E., Glenn J., Rosolowsky E., 2018, MNRAS, 480, 893

Zhang Z.-Y. e. a., 2014, ApJL, 784, L31

Zimmermann T., Stutzki J., 1992, Physica A, 191, 79